Information Retrieval

Oct 1, 2024 @ Introduction to Human Language Technology

Eugene Yang eugene.yang@jhu.edu

Slides borrowed from SIGIR24 Tutorial
“Neural Methods for Cross-Language Information Retrieval”

What is Information Retrieval?

(relevant)

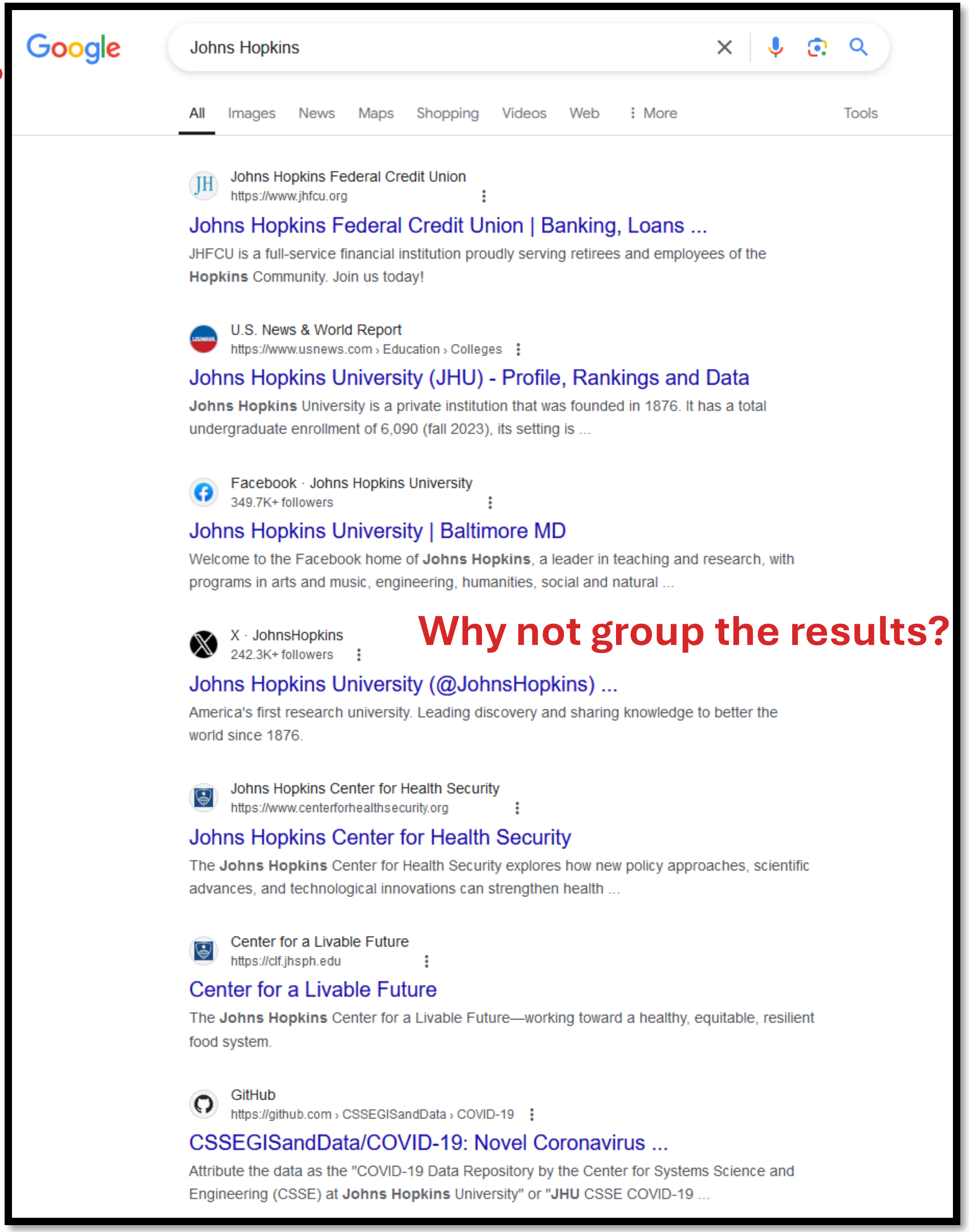
Retrieve information from a storage based on user's information need

Don’t we have Google?

Yes, **but Google is not all.**

What if I’m looking for the person?

Why in a list?



Why not group the results?

Why not read my mind?

Why asking me to read?

Google Search is just one implementation

Google trained us well!

- Even faster?
- Smarter?
- Cross language?

Hard Matching Problem

- Text to text
  - Search in notes
  - Cross language search
  - Cross domain search
- Text to other modalities
  - Image search
  - Video search

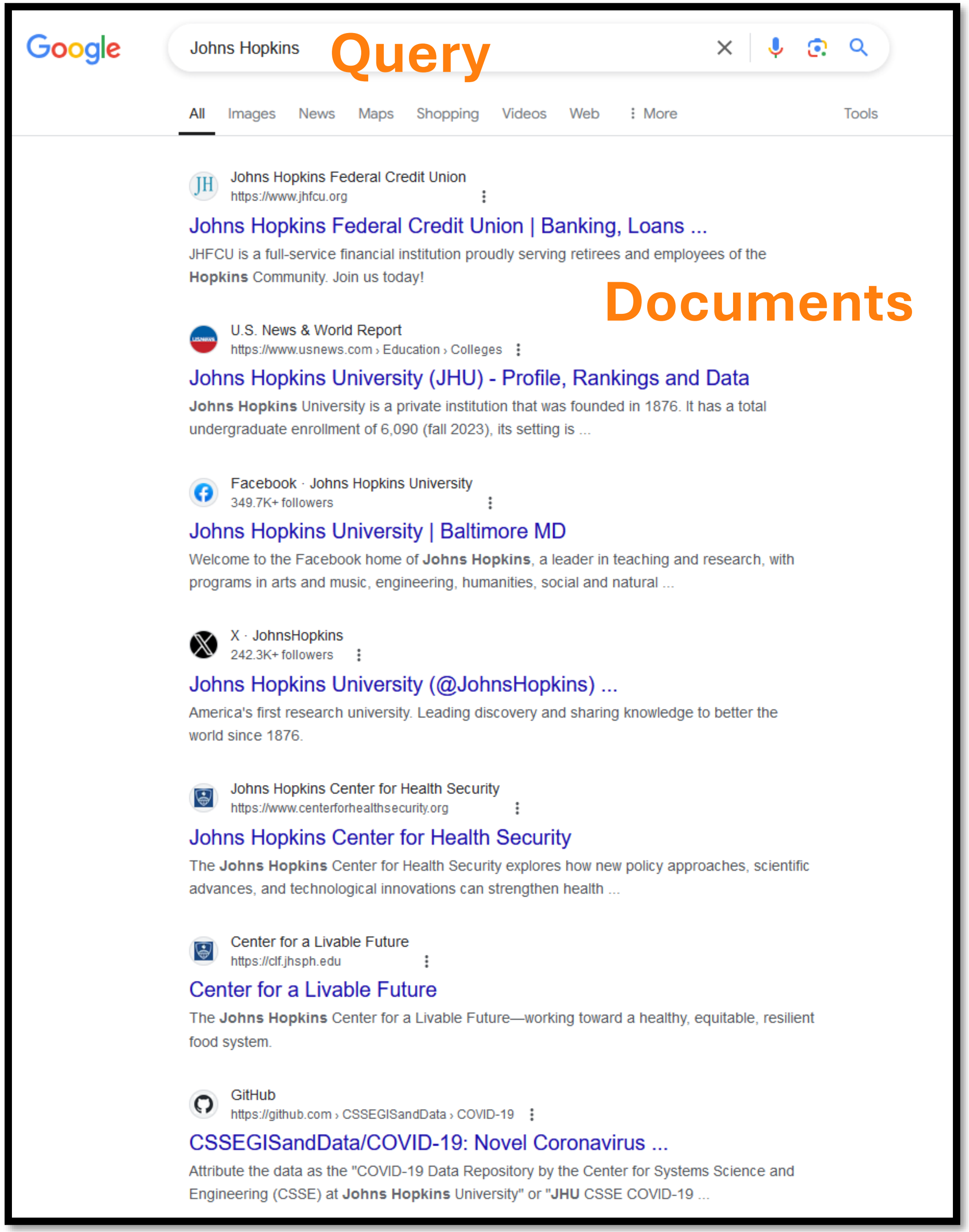
Different Search Process

- Iterative search
  - e.g., electronic discovery and systematic review
- Conversational search
  - Alexa search
- Recommendation systems
- (Set Retrieval)

Core Problem

- Matching problem
- Do it fast

Ranked List



Agenda

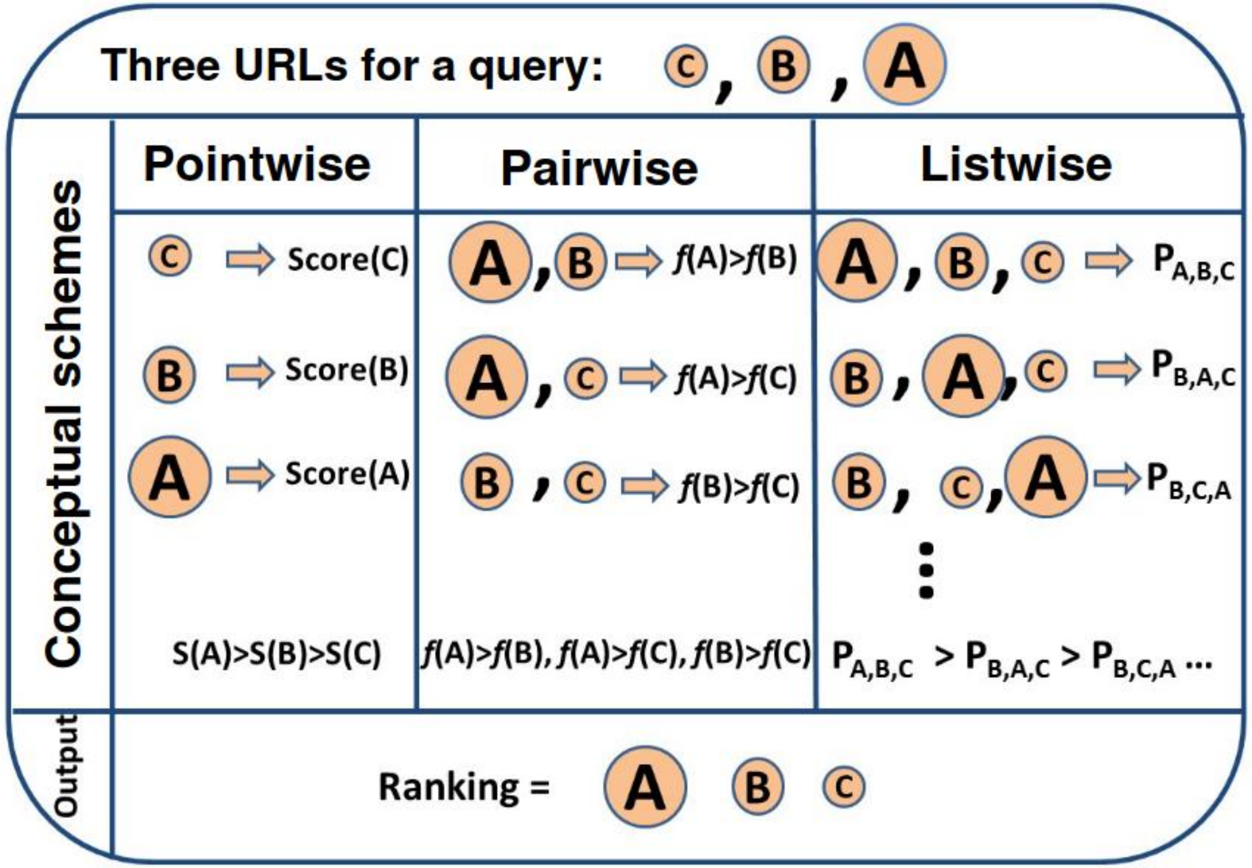
- What is information retrieval?
- Retrieval Modeling and Pipeline
  - Statistical and Neural
- Evaluation
- State of IR Research and active research problems

Retrieval Modeling and Pipeline

Three main ways

- **Pointwise**
- Pairwise
- Listwise

- And combinations of them



https://medium.com/vptech/learning-to-rank-at-veepee-ed420fd828e5

Statistical Models

$$\text{score}(D, Q) = \sum_{\text{For each query term}} \boxed{\text{How important the term is}} \times \boxed{\text{How often the term appear in the D}}$$

$$\text{score}(D, Q) = \sum_{\text{For each query term}} \boxed{\text{Inverted document frequency}} \times \boxed{\text{Term frequency}}$$

TF-iDF

$$\text{score}(D, Q) = \sum_{i=1}^{n} log\frac{N}{n_t} \times \log(f(q_i, D) + 1)$$

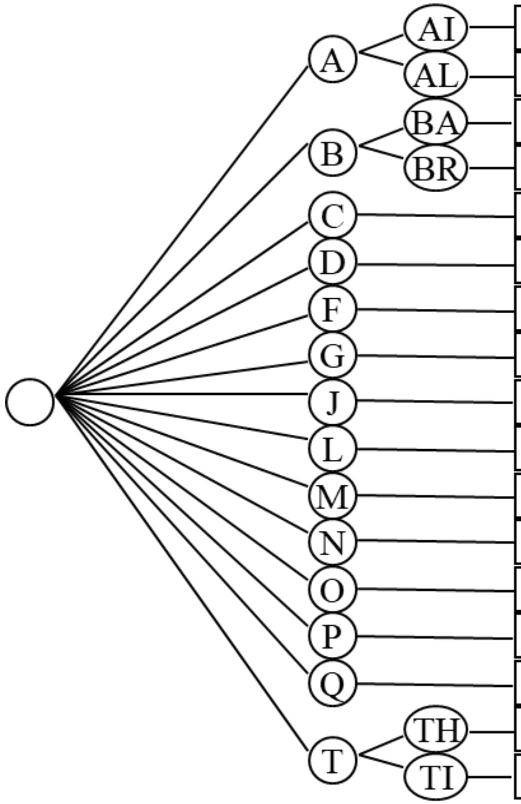
BM25

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

How to make it fast?

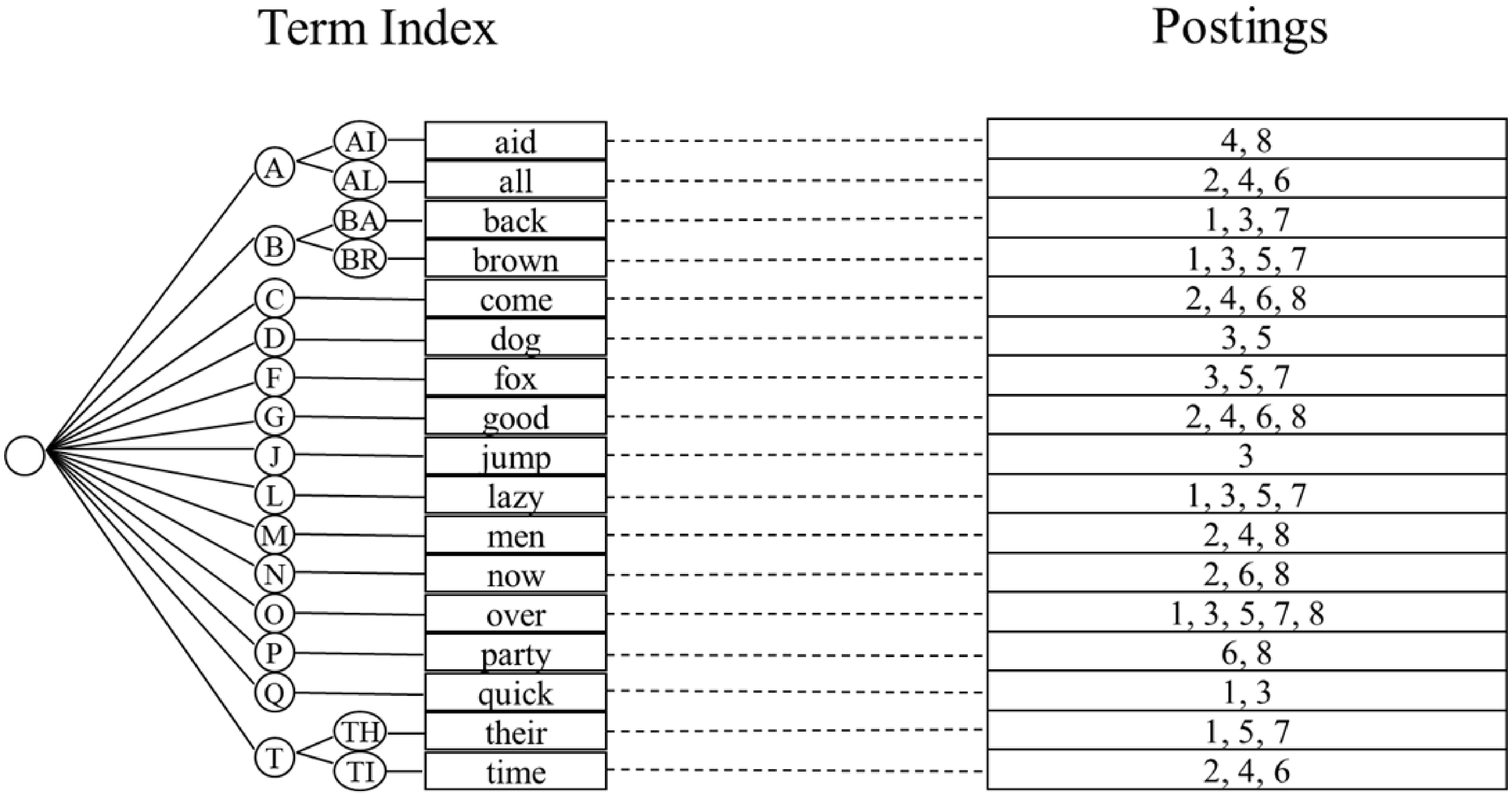
- “Fast” in responding to queries
- Better data structure
- Preprocess the data

Inverted Index

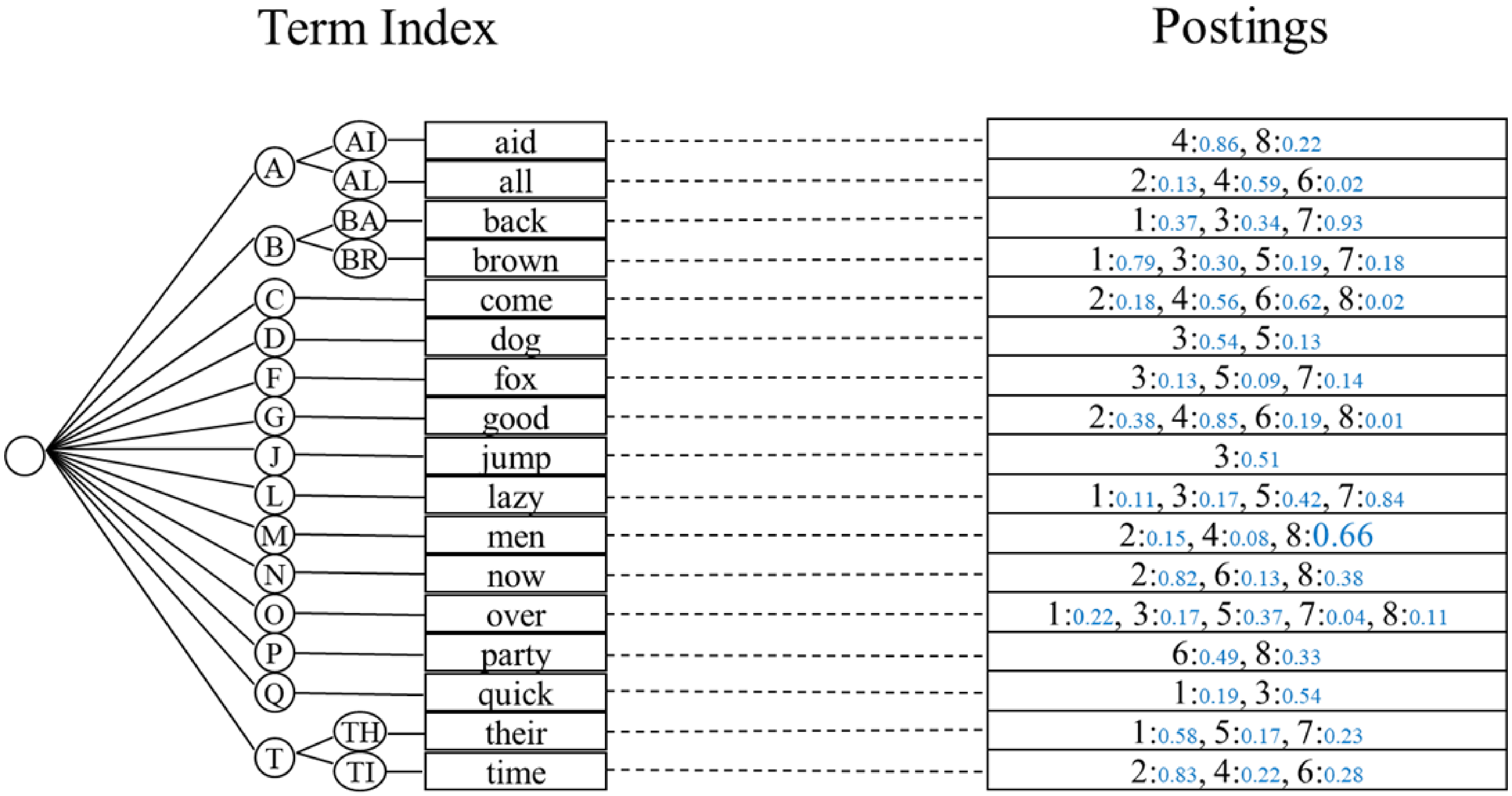


Term Index	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Postings
aid				1				1	4, 8
all		1		1		1			2, 4, 6
back	1		1				1		1, 3, 7
brown	1		1		1		1		1, 3, 5, 7
come		1		1		1		1	2, 4, 6, 8
dog			1		1				3, 5
fox			1		1		1		3, 5, 7
good		1		1		1		1	2, 4, 6, 8
jump			1						3
lazy	1		1		1		1		1, 3, 5, 7
men		1		1				1	2, 4, 8
now		1				1		1	2, 6, 8
over	1		1		1		1	1	1, 3, 5, 7, 8
party						1		1	6, 8
quick	1		1						1, 3
their	1				1		1		1, 5, 7
time		1		1		1			2, 4, 6

Inverted Index

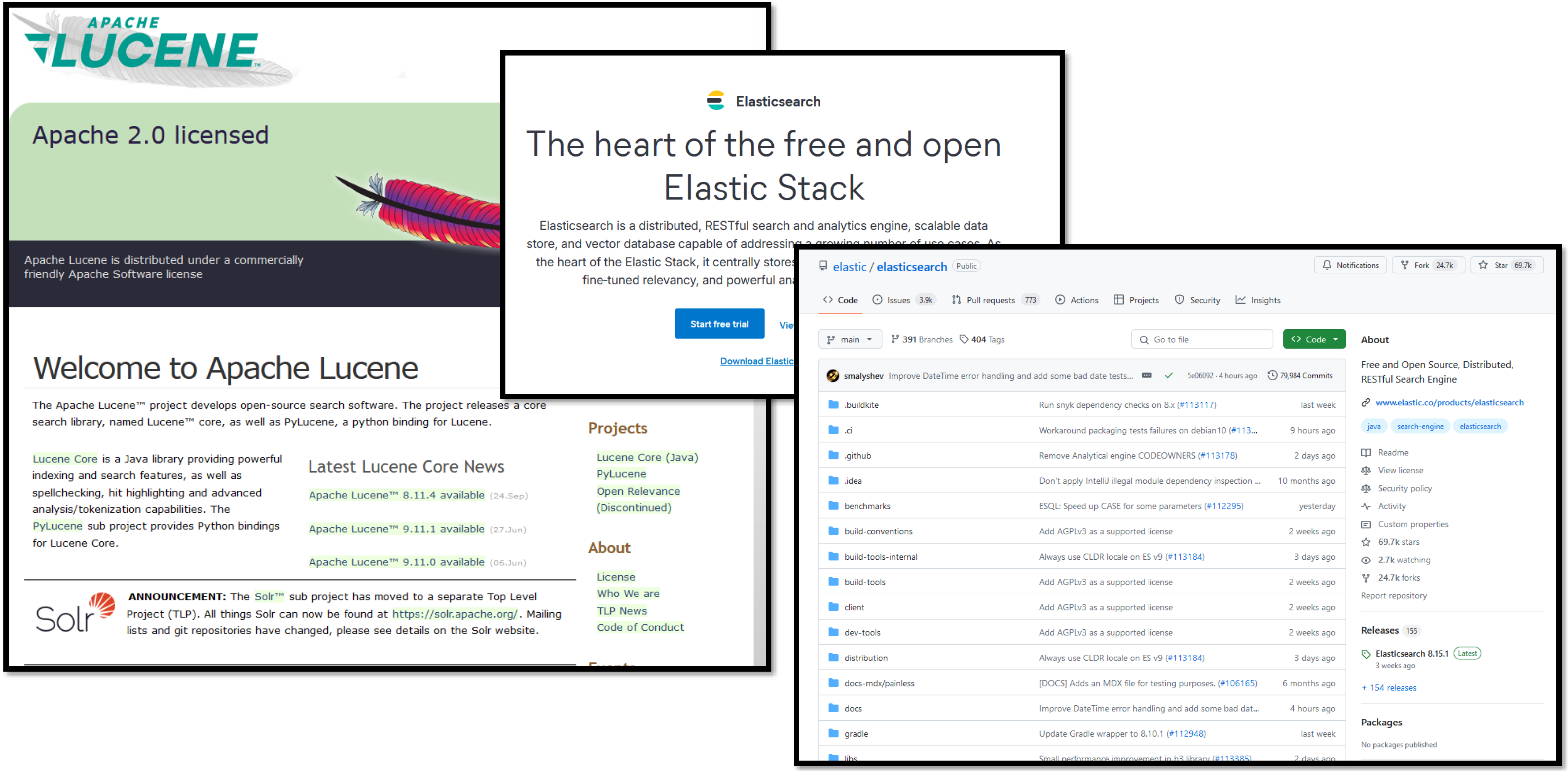


Inverted Index



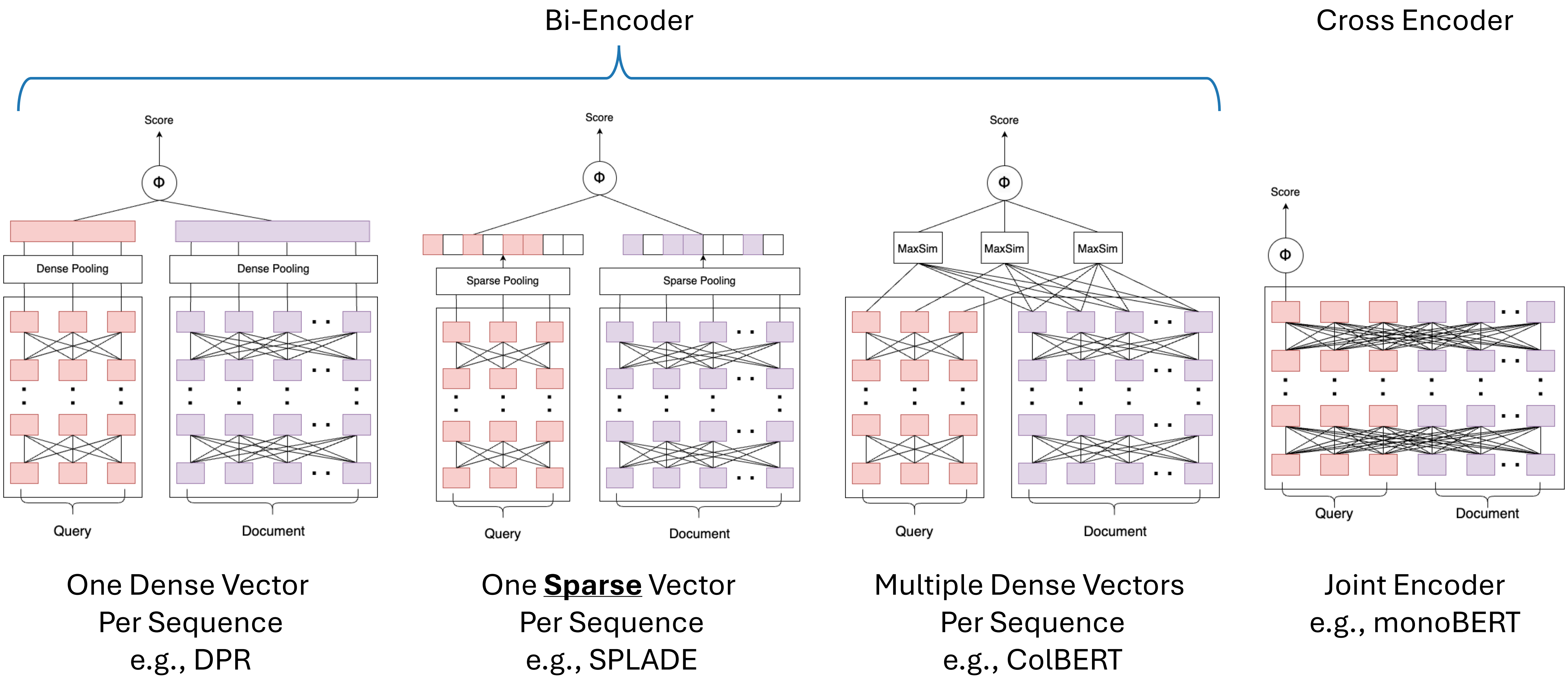
Two-Stage System

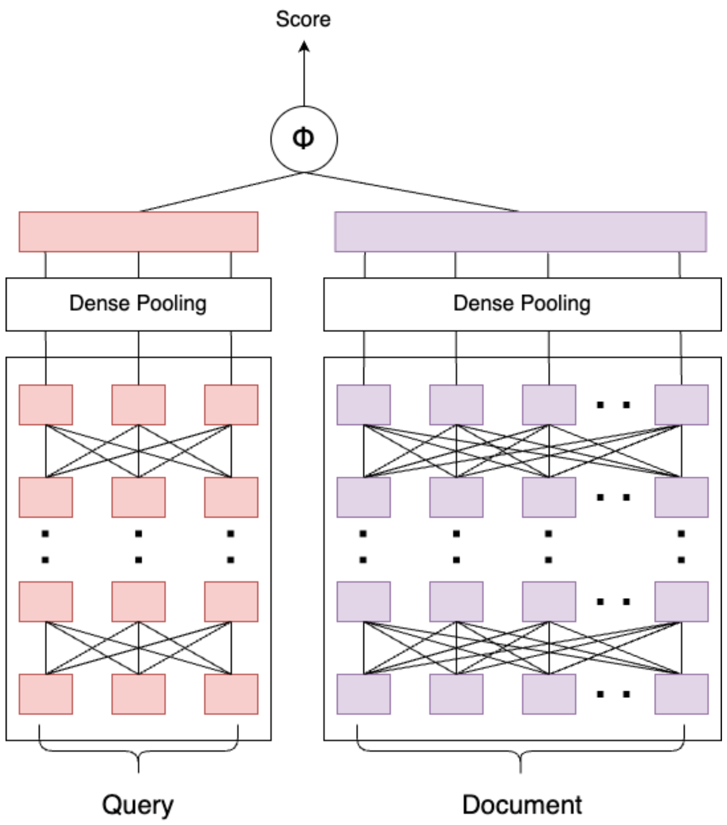
- Offline preprocessing and indexing
  - Build the inverted index
- Online query serving
  - Traverse the inverted index and score it


APACHE LUCENE
Apache 2.0 licensed
Apache Lucene is distributed under a commercially friendly Apache Software license
Welcome to Apache Lucene
The Apache Lucene™ project develops open-source search software. The project releases a core search library, named Lucene™ core, as well as PyLucene, a python binding for Lucene.
Lucene Core is a Java library providing powerful indexing and search features, as well as spellchecking, hit highlighting and advanced analysis/tokenization capabilities. The PyLucene sub project provides Python bindings for Lucene Core.
Latest Lucene Core News
Apache Lucene™ 8.11.4 available (24.Sep)
Apache Lucene™ 9.11.1 available (27.Jun)
Apache Lucene™ 9.11.0 available (06.Jun)
Projects
Lucene Core (Java)
PyLucene
Open Relevance (Discontinued)
About
License
Who We are
TLP News
Code of Conduct
Solr
ANNOUNCEMENT: The Solr™ sub project has moved to a separate Top Level Project (TLP). All things Solr can now be found at https://solr.apache.org/. Mailing lists and git repositories have changed, please see details on the Solr website.
Elasticsearch
The heart of the free and open Elastic Stack
Elasticsearch is a distributed, RESTful search and analytics engine, scalable data store, and vector database capable of addressing
the heart of the Elastic Stack, it centrally stores
fine-tuned relevancy, and powerful an
Start free trial
Download Elastic
elastic / elasticsearch
Public
Notifications
Fork 24.7k
Star 69.7k
Code
Issues 3.9k
Pull requests 773
Actions
Projects
Security
Insights
main
391 Branches
404 Tags
Go to file
Code
smalyshev Improve DateTime error handling and add some bad date tests...
5e06092 · 4 hours ago
79,984 Commits
.buildkite Run snyk dependency checks on 8.x (#113117) last week
.ci Workaround packaging tests failures on debian10 (#113... 9 hours ago
.github Remove Analytical engine CODEOWNERS (#113178) 2 days ago
.idea Don't apply IntelliJ illegal module dependency inspection ... 10 months ago
benchmarks ESQL: Speed up CASE for some parameters (#112295) yesterday
build-conventions Add AGPLv3 as a supported license 2 weeks ago
build-tools-internal Always use CLDR locale on ES v9 (#113184) 3 days ago
build-tools Add AGPLv3 as a supported license 2 weeks ago
client Add AGPLv3 as a supported license 2 weeks ago
dev-tools Add AGPLv3 as a supported license 2 weeks ago
distribution Always use CLDR locale on ES v9 (#113184) 3 days ago
docs-mdx/painless [DOCS] Adds an MDX file for testing purposes. (#106165) 6 months ago
docs Improve DateTime error handling and add some bad dat... 4 hours ago
gradle Update Gradle wrapper to 8.10.1 (#112948) last week
About
Free and Open Source, Distributed, RESTful Search Engine
www.elastic.co/products/elasticsearch
java
search-engine
elasticsearch
Readme
View license
Security policy
Activity
Custom properties
69.7k stars
2.7k watching
24.7k forks
Report repository
Releases 155
Elasticsearch 8.15.1 Latest
3 weeks ago
+ 154 releases
Packages
No packages published

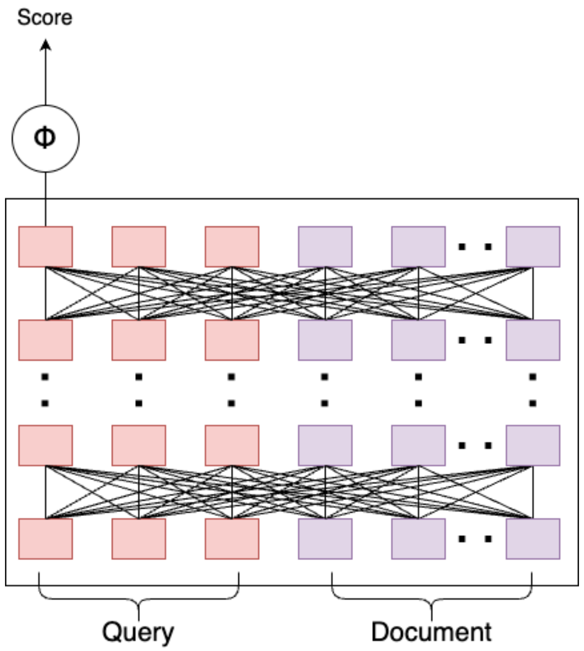
Can we go beyond surface forms?

neural language models


Bi-Encoder
Cross Encoder
Score
Φ
Dense Pooling
Dense Pooling
Query
Document
Score
Φ
Sparse Pooling
Sparse Pooling
Query
Document
Score
Φ
MaxSim
MaxSim
MaxSim
Query
Document
Score
Φ
Query
Document
One Dense Vector
Per Sequence
e.g., DPR
One Sparse Vector
Per Sequence
e.g., SPLADE
Multiple Dense Vectors
Per Sequence
e.g., ColBERT
Joint Encoder
e.g., monoBERT



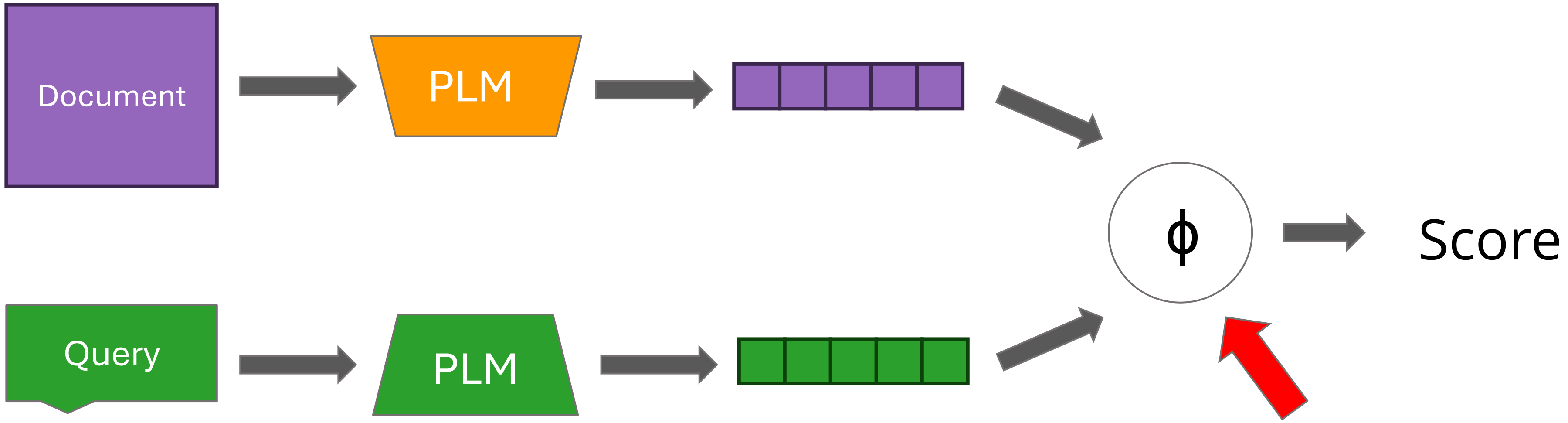
Online Offline

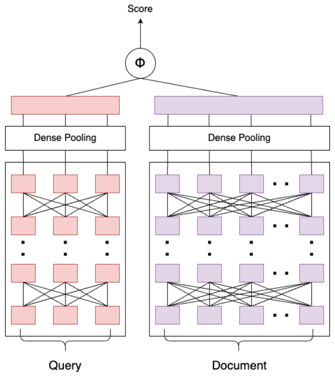


Both Online

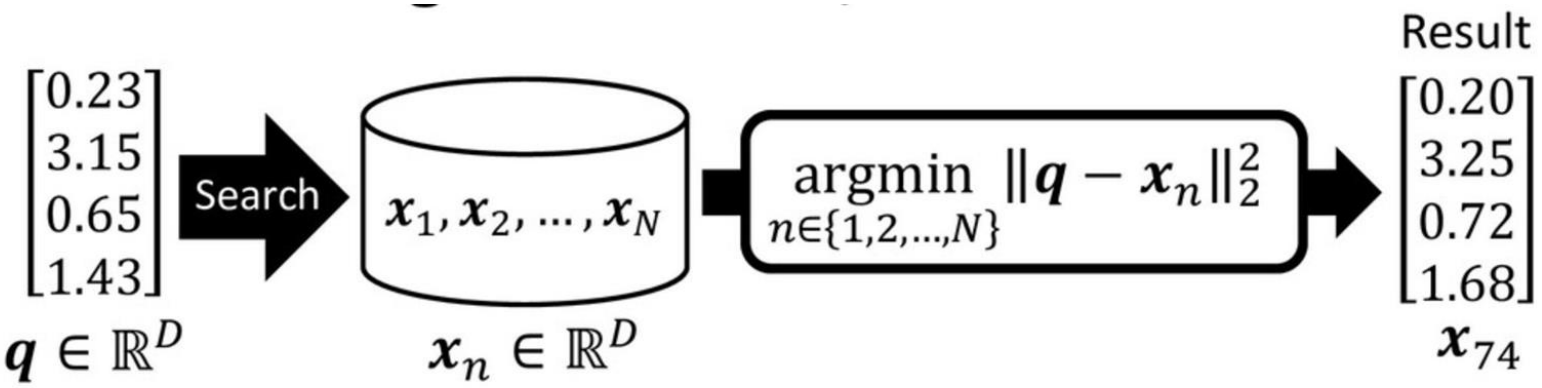
Separate query and document processing

One Vector per Query, One Vector per Document





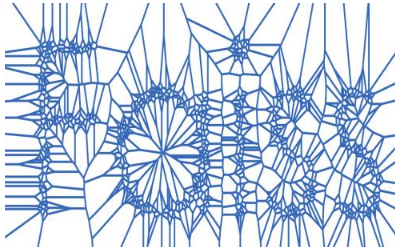
Nearest Vectors aka Neighbors



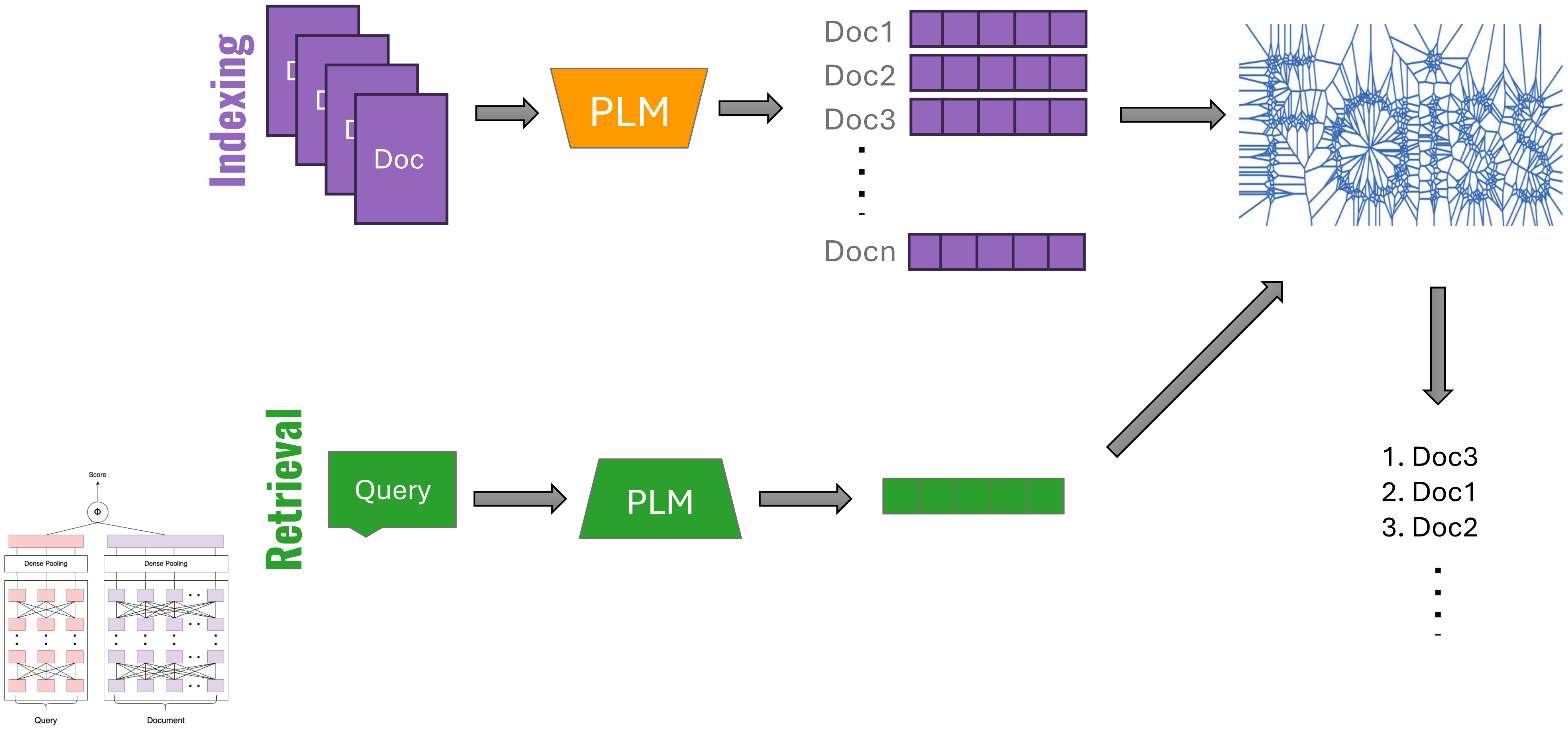
- Linear Search
  - Slow (scales linearly in size of document collection)

- Approximate Methods (e.g., Product Quantization) → **ANN**
  - Faster Search

- Runtime Efficiency vs Effectiveness

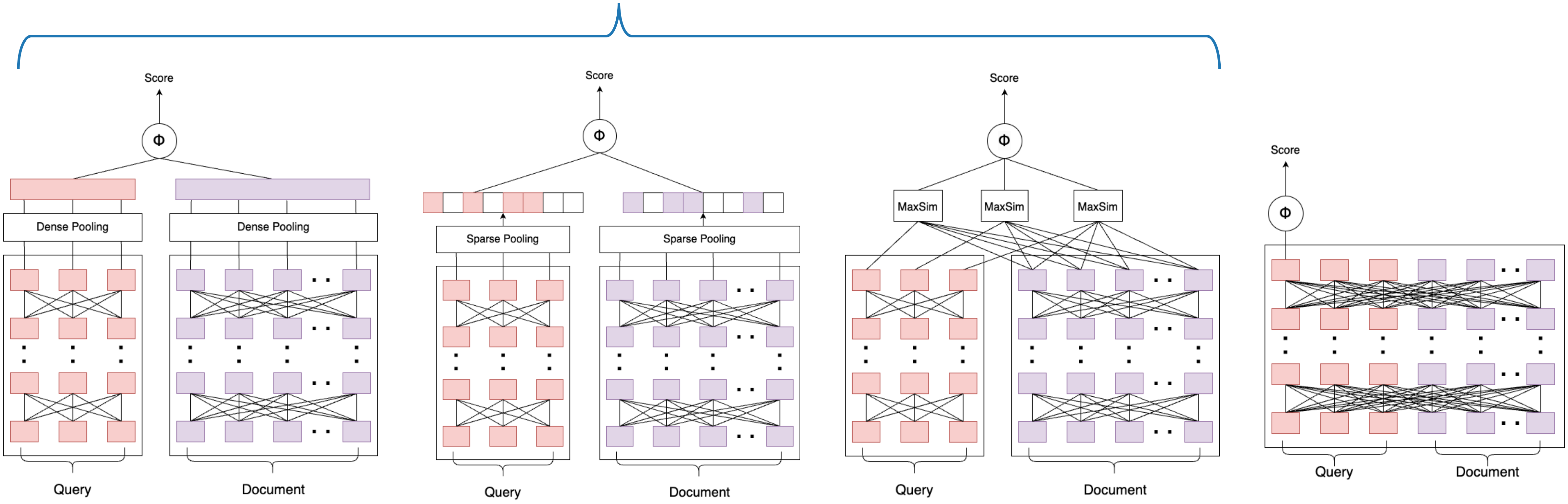


DPR Indexing and Retrieval



Bi-Encoder

Cross Encoder



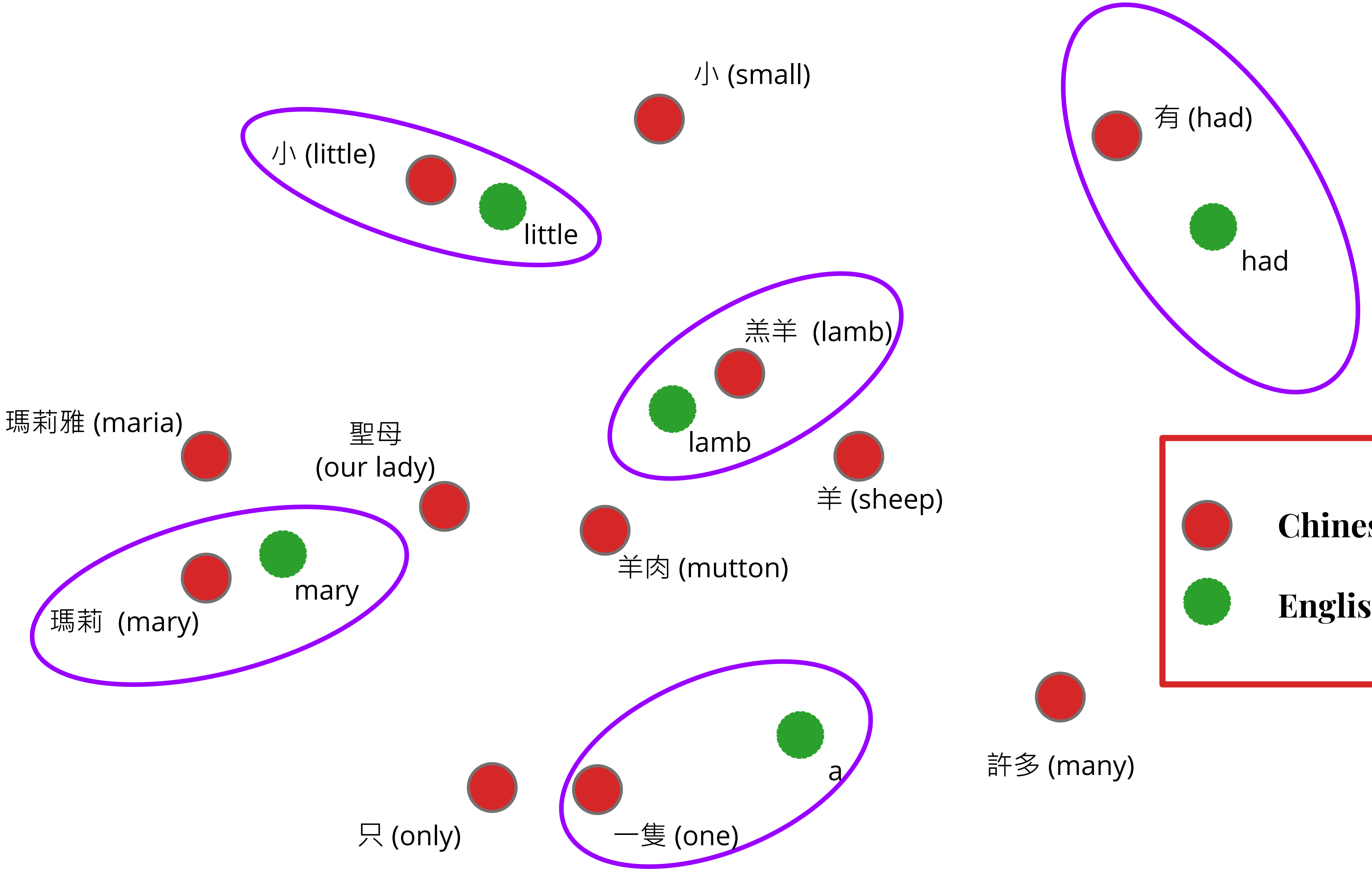
One Dense Vector
Per Sequence
e.g., DPR

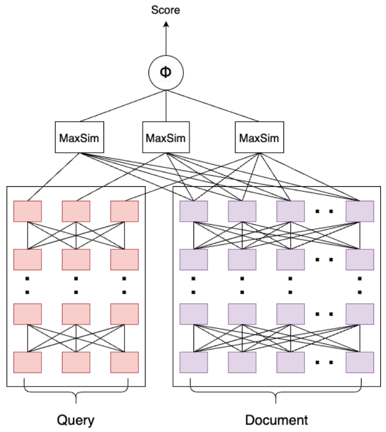
One **Sparse** Vector
Per Sequence
e.g., SPLADE

Multiple Dense Vectors
Per Sequence
e.g., ColBERT

Joint Encoder
e.g., monoBERT

One Vector per Term: MaxSim

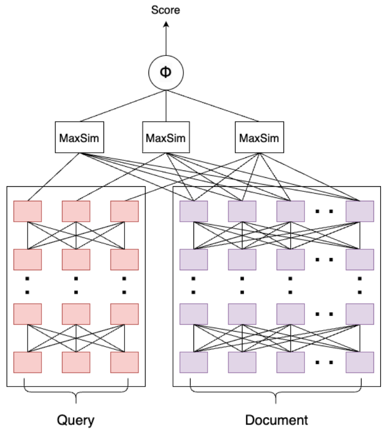


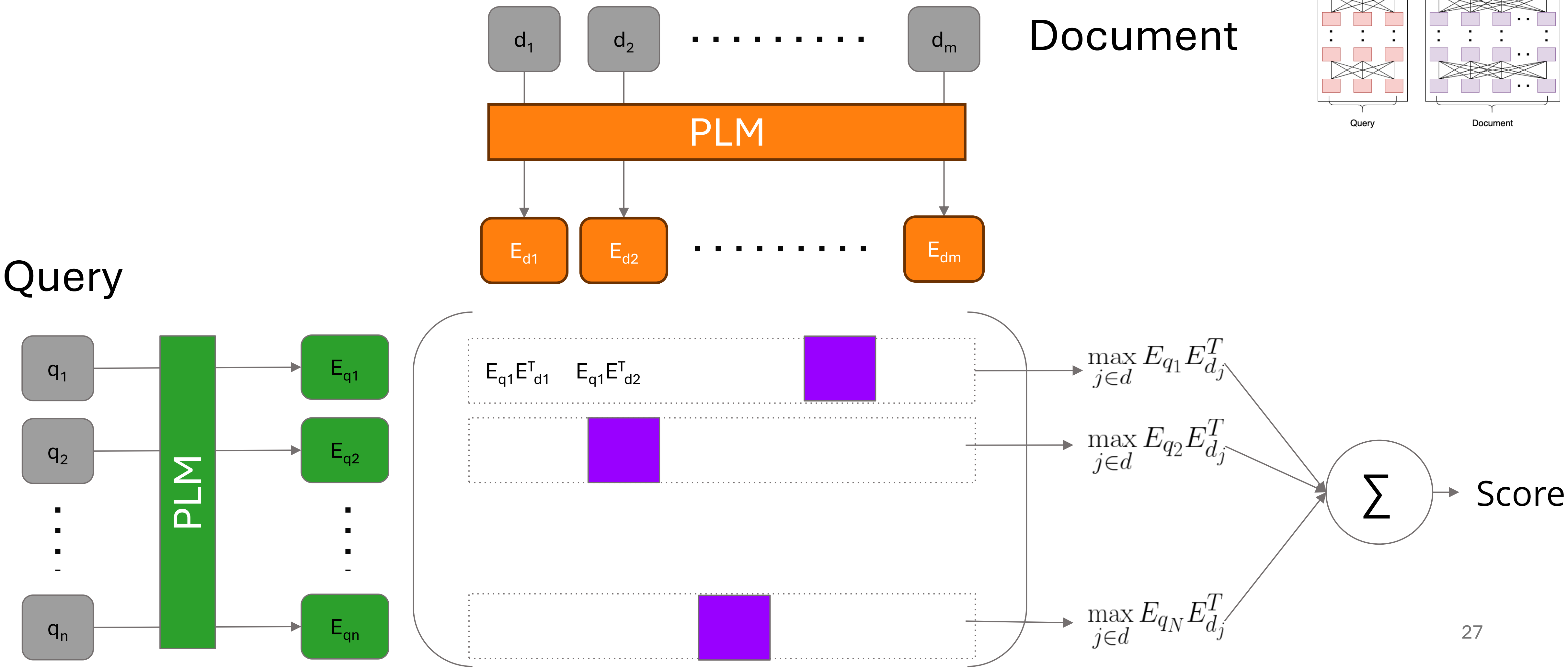


Chinese Document term embedding

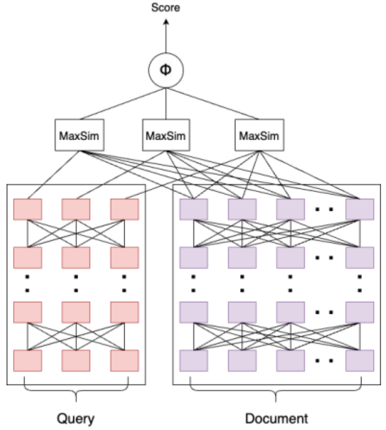
English Query term embedding

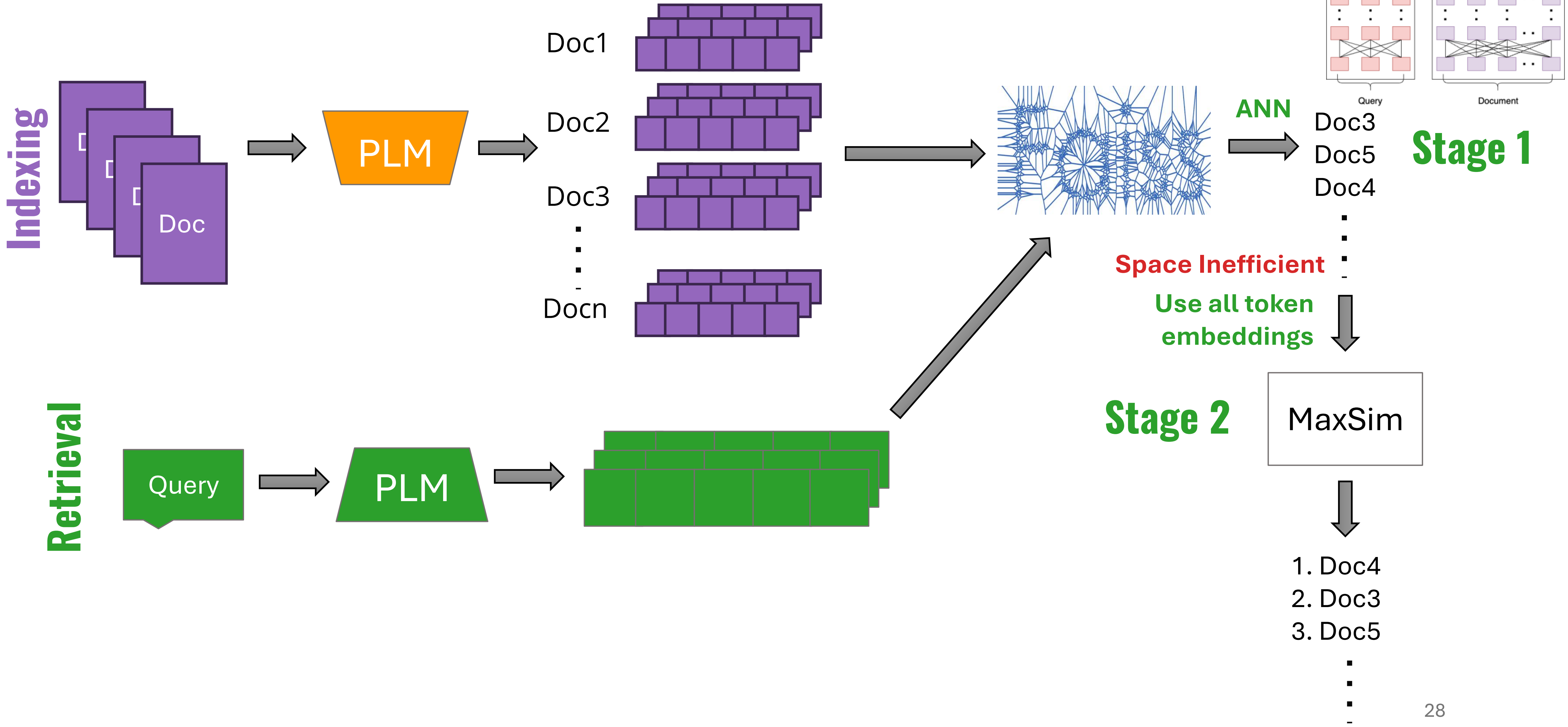
MaxSim in Action -- ColBERT



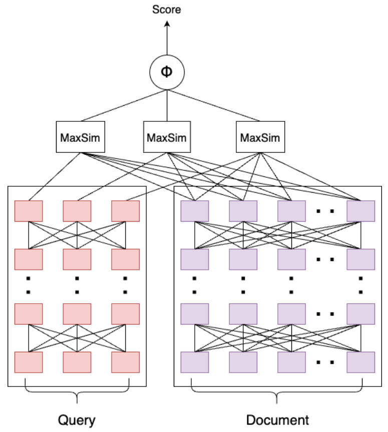


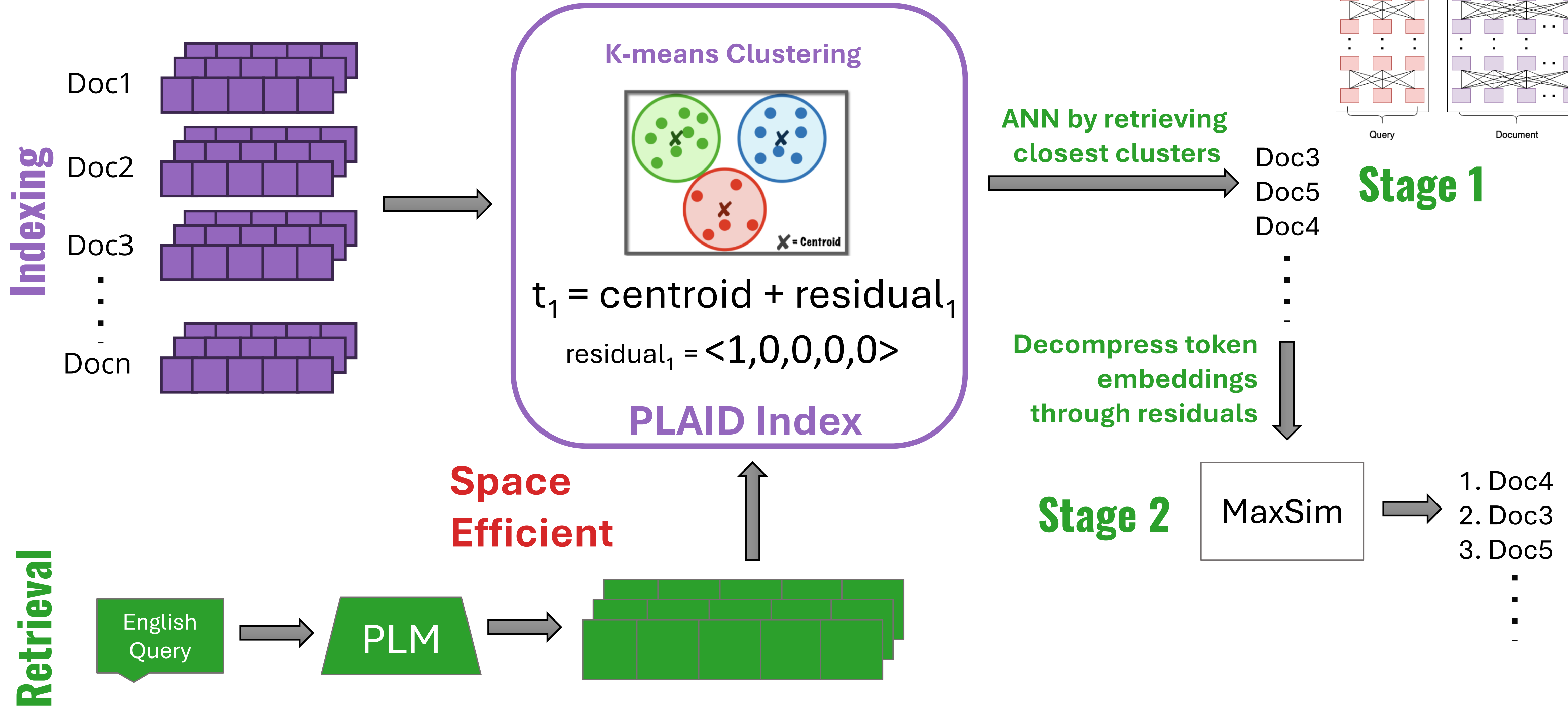
One Vector per term: Multi-stage Retrieval

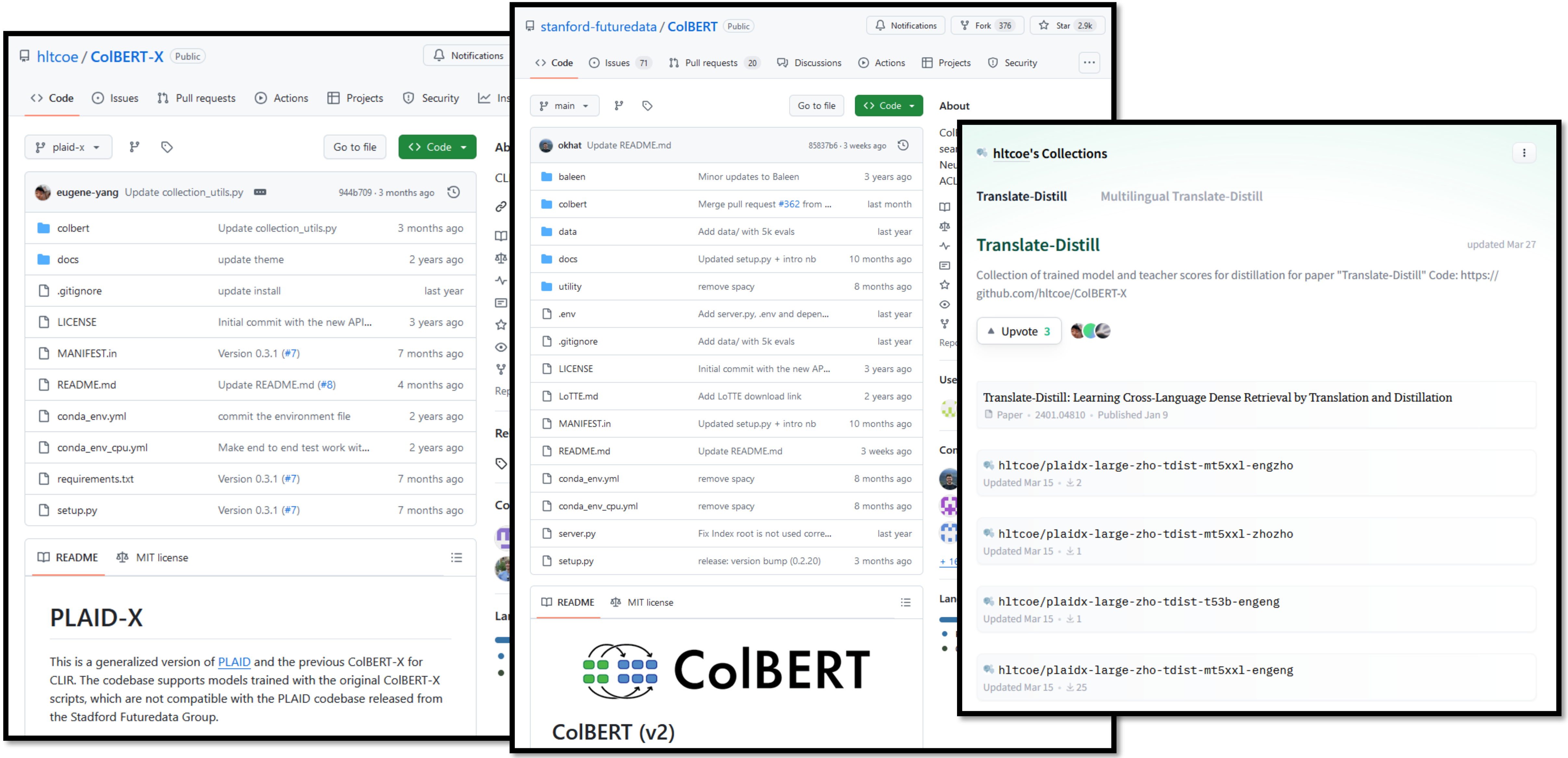




Efficient PLAID Indexing Architecture

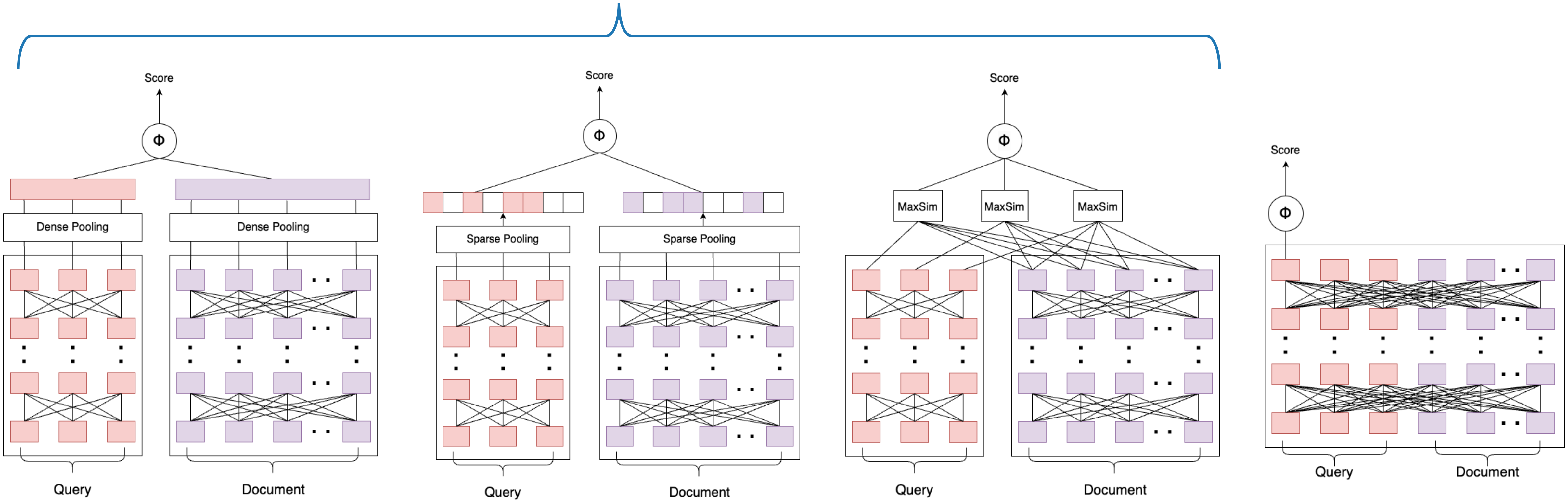





hltcoe / ColBERT-X Public
Notifications
Code Issues Pull requests Actions Projects Security
plaid-x
Go to file
Code
eugene-yang Update collection_utils.py 944b709 · 3 months ago
colbert Update collection_utils.py 3 months ago
docs update theme 2 years ago
.gitignore update install last year
LICENSE Initial commit with the new API... 3 years ago
MANIFEST.in Version 0.3.1 (#7) 7 months ago
README.md Update README.md (#8) 4 months ago
conda_env.yml commit the environment file 2 years ago
conda_env_cpu.yml Make end to end test work wit... 2 years ago
requirements.txt Version 0.3.1 (#7) 7 months ago
setup.py Version 0.3.1 (#7) 7 months ago
README MIT license
PLAID-X
This is a generalized version of PLAID and the previous ColBERT-X for CLIR. The codebase supports models trained with the original ColBERT-X scripts, which are not compatible with the PLAID codebase released from the Stadford Futuredata Group.
stanford-futuredata / ColBERT Public
Notifications Fork 376 Star 2.9k
Code Issues 71 Pull requests 20 Discussions Actions Projects Security
main
Go to file
Code
About
okhat Update README.md 85837b6 · 3 weeks ago
baleen Minor updates to Baleen 3 years ago
colbert Merge pull request #362 from ... last month
data Add data/ with 5k evals last year
docs Updated setup.py + intro nb 10 months ago
utility remove spacy 8 months ago
.env Add server.py, .env and depen... last year
.gitignore Add data/ with 5k evals last year
LICENSE Initial commit with the new AP... 3 years ago
LoTTE.md Add LoTTE download link 2 years ago
MANIFEST.in Updated setup.py + intro nb 10 months ago
README.md Update README.md 3 weeks ago
conda_env.yml remove spacy 8 months ago
conda_env_cpu.yml remove spacy 8 months ago
server.py Fix Index root is not used corre... last year
setup.py release: version bump (0.2.20) 3 months ago
README MIT license
ColBERT
ColBERT (v2)
hltcoe's Collections
Translate-Distill
Multilingual Translate-Distill
Translate-Distill
updated Mar 27
Collection of trained model and teacher scores for distillation for paper "Translate-Distill" Code: https://github.com/hltcoe/ColBERT-X
Upvote 3
Translate-Distill: Learning Cross-Language Dense Retrieval by Translation and Distillation
Paper · 2401.04810 · Published Jan 9
hltcoe/plaidx-large-zho-tdist-mt5xxl-engzho
Updated Mar 15 · 2
hltcoe/plaidx-large-zho-tdist-mt5xxl-zhozho
Updated Mar 15 · 1
hltcoe/plaidx-large-zho-tdist-t53b-engeng
Updated Mar 15 · 1
hltcoe/plaidx-large-zho-tdist-mt5xxl-engeng
Updated Mar 15 · 25

Bi-Encoder

Cross Encoder



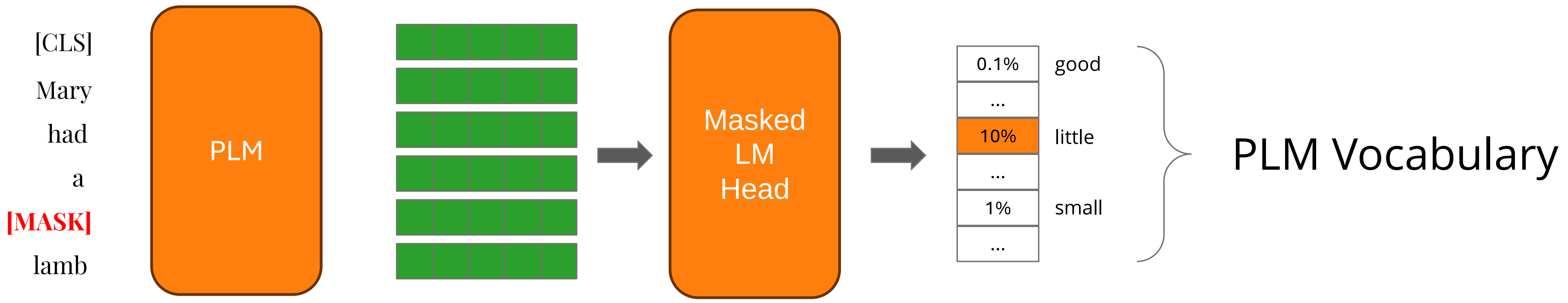
One Dense Vector
Per Sequence
e.g., DPR

One **Sparse** Vector
Per Sequence
e.g., SPLADE

Multiple Dense Vectors
Per Sequence
e.g., ColBERT

Joint Encoder
e.g., monoBERT

High-dimensional Vector: Masked LM



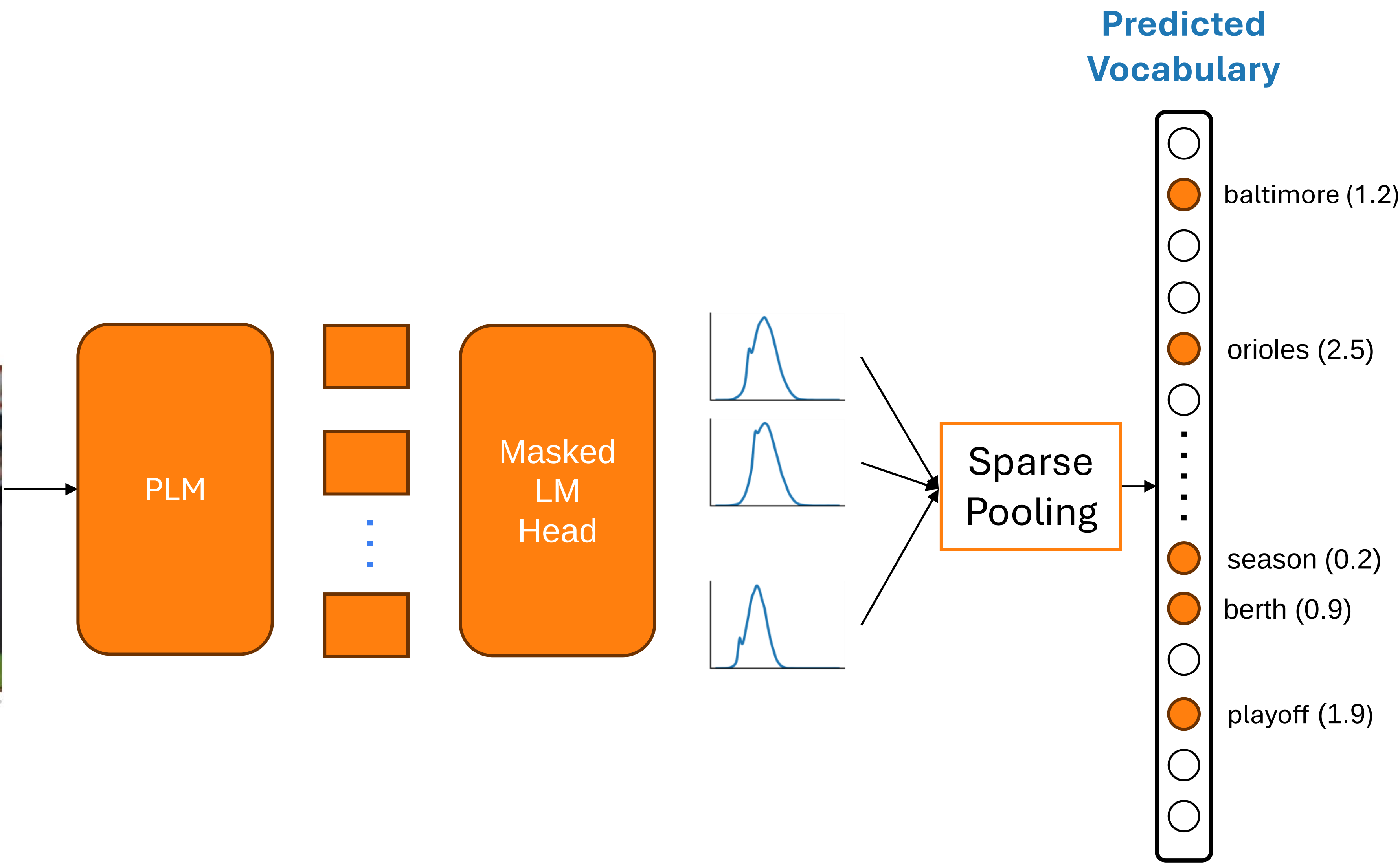
SPLADE

Baltimore Orioles clinch playoff berth for 2nd straight season

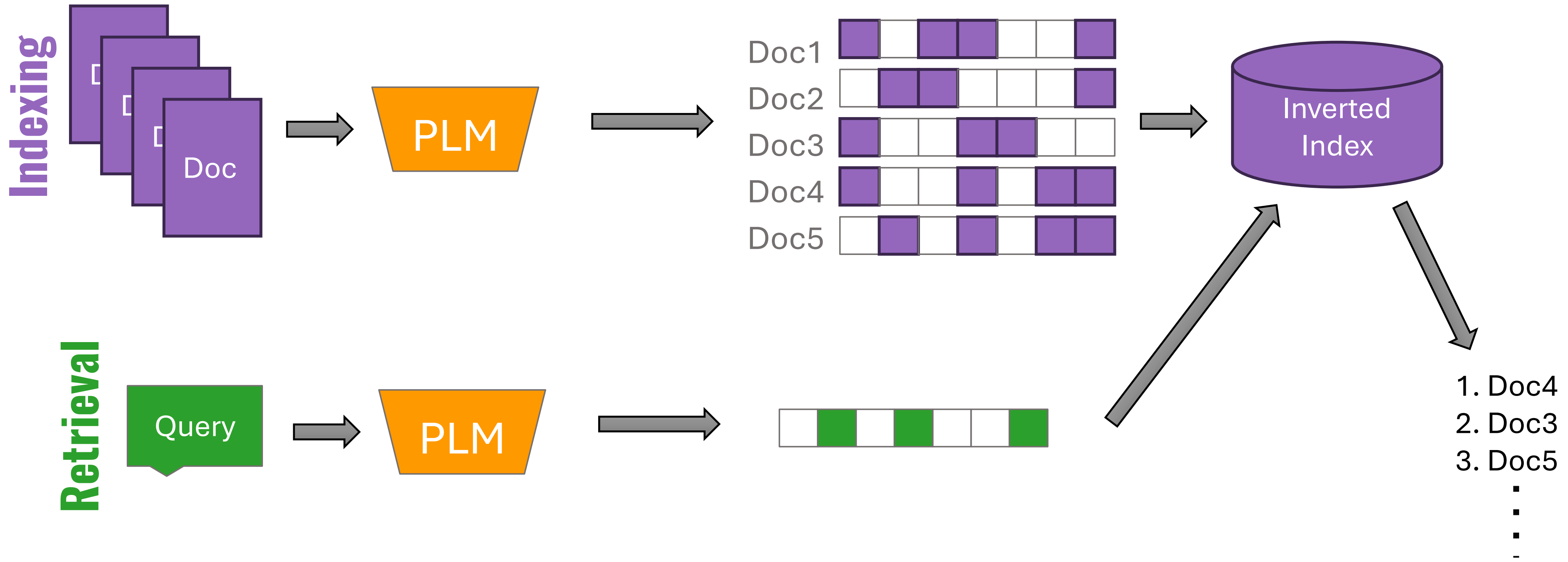
The Baltimore Orioles are headed to the playoffs in consecutive years for the first time since the 1990s, clinching no worse than a wild-card berth with a 5-3 win over the New York Yankees paired with Minnesota's loss to Miami

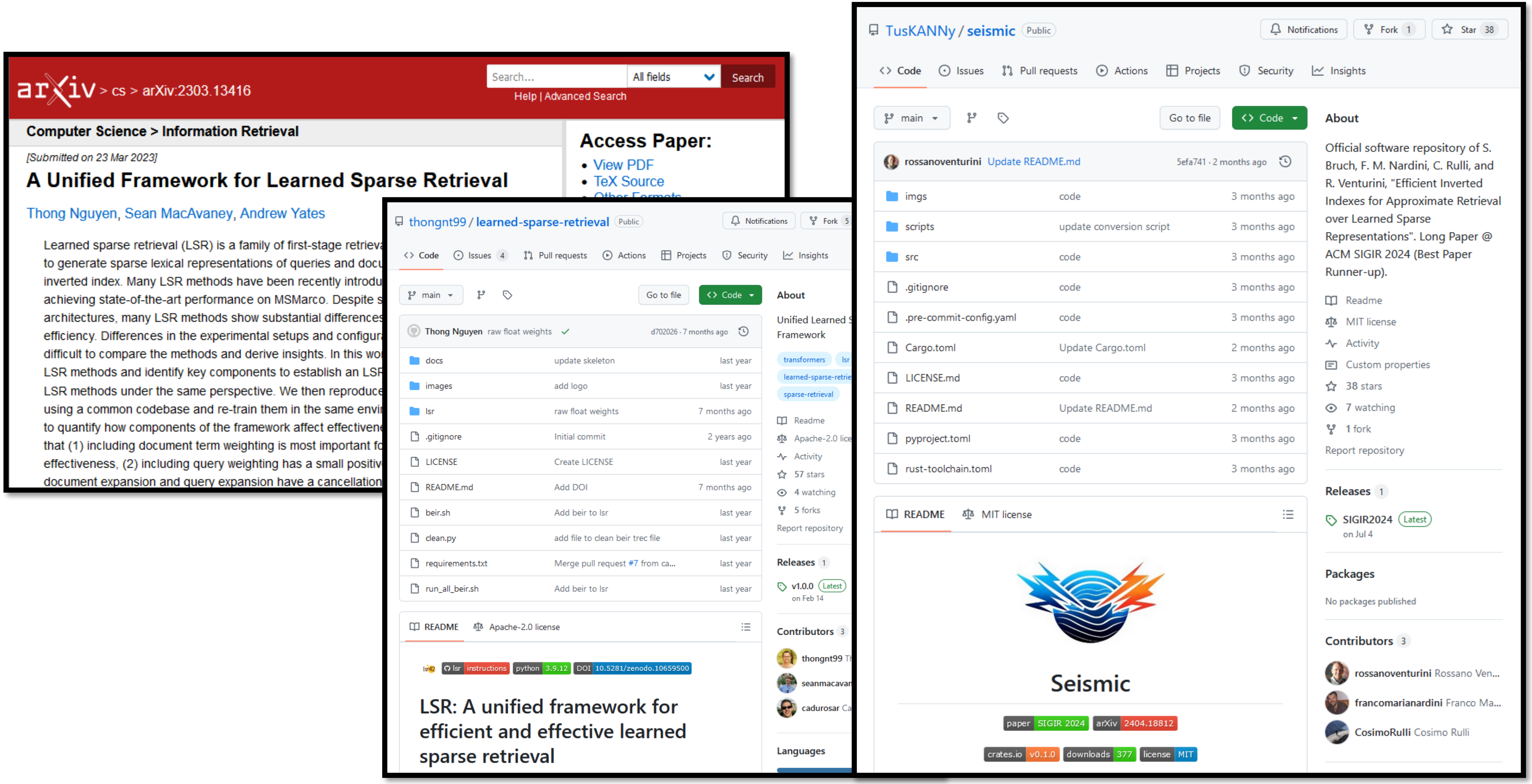
Credit: AP

Baltimore Orioles' Anthony Santander runs the bases after hitting a home run during the sixth inning of a baseball game against the New York Yankees, Tuesday, Sept. 24, 2024, in New York. (AP Photo/Bryan Woolston)



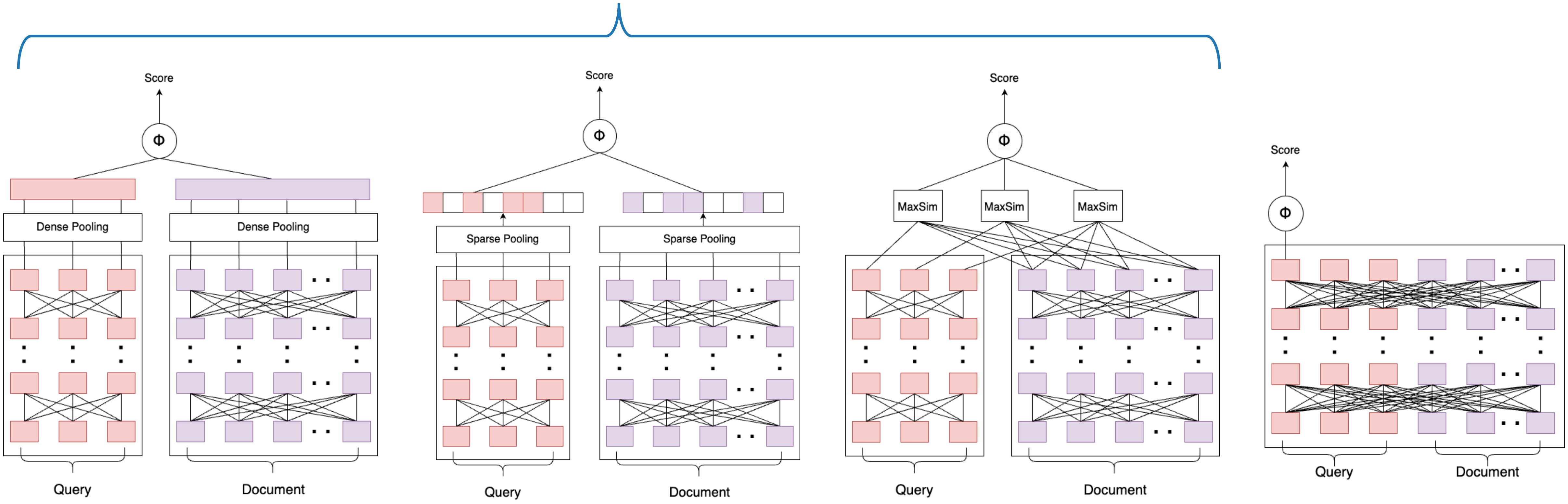
SPLADE Search Pipeline




arXiv > cs > arXiv:2303.13416
Search...
All fields
Search
Help | Advanced Search
Computer Science > Information Retrieval
[Submitted on 23 Mar 2023]
A Unified Framework for Learned Sparse Retrieval
Thong Nguyen, Sean MacAvaney, Andrew Yates
Learned sparse retrieval (LSR) is a family of first-stage retrieva
to generate sparse lexical representations of queries and docu
inverted index. Many LSR methods have been recently introdu
achieving state-of-the-art performance on MSMarco. Despite s
architectures, many LSR methods show substantial differences
efficiency. Differences in the experimental setups and configura
difficult to compare the methods and derive insights. In this wo
LSR methods and identify key components to establish an LSR
LSR methods under the same perspective. We then reproduce
using a common codebase and re-train them in the same envir
to quantify how components of the framework affect effectivene
that (1) including document term weighting is most important fo
effectiveness, (2) including query weighting has a small positiv
document expansion and query expansion have a cancellation
Access Paper:
View PDF
TeX Source
thongnt99 / learned-sparse-retrieval Public
Notifications
Fork 5
Code
Issues 4
Pull requests
Actions
Projects
Security
Insights
main
Go to file
Code
Thong Nguyen raw float weights
d702026 · 7 months ago
docs update skeleton last year
images add logo last year
lsr raw float weights 7 months ago
.gitignore Initial commit 2 years ago
LICENSE Create LICENSE last year
README.md Add DOI 7 months ago
beir.sh Add beir to lsr last year
clean.py add file to clean beir trec file last year
requirements.txt Merge pull request #7 from ca... last year
run_all_beir.sh Add beir to lsr last year
README
Apache-2.0 license
lsr instructions
python 3.9.12
DOI 10.5281/zenodo.10659500
LSR: A unified framework for efficient and effective learned sparse retrieval
About
Unified Learned
Framework
transformers
lsr
learned-sparse-retrie
sparse-retrieval
Readme
Apache-2.0 lice
Activity
57 stars
4 watching
5 forks
Report repository
Releases 1
v1.0.0 Latest
on Feb 14
Contributors 3
thongnt99
seanmacavan
cadurosar
Languages
TusKANNy / seismic Public
Notifications
Fork 1
Star 38
Code
Issues
Pull requests
Actions
Projects
Security
Insights
main
Go to file
Code
rossanoventurini Update README.md
5efa741 · 2 months ago
imgs code 3 months ago
scripts update conversion script 3 months ago
src code 3 months ago
.gitignore code 3 months ago
.pre-commit-config.yaml code 3 months ago
Cargo.toml Update Cargo.toml 2 months ago
LICENSE.md code 3 months ago
README.md Update README.md 2 months ago
pyproject.toml code 3 months ago
rust-toolchain.toml code 3 months ago
README
MIT license
Seismic
paper SIGIR 2024
arXiv 2404.18812
crates.io v0.1.0
downloads 377
license MIT
About
Official software repository of S. Bruch, F. M. Nardini, C. Rulli, and R. Venturini, "Efficient Inverted Indexes for Approximate Retrieval over Learned Sparse Representations". Long Paper @ ACM SIGIR 2024 (Best Paper Runner-up).
Readme
MIT license
Activity
Custom properties
38 stars
7 watching
1 fork
Report repository
Releases 1
SIGIR2024 Latest
on Jul 4
Packages
No packages published
Contributors 3
rossanoventurini Rossano Ven...
francomarianardini Franco Ma...
CosimoRulli Cosimo Rulli

Bi-Encoder

Cross Encoder



One Dense Vector
Per Sequence
e.g., DPR

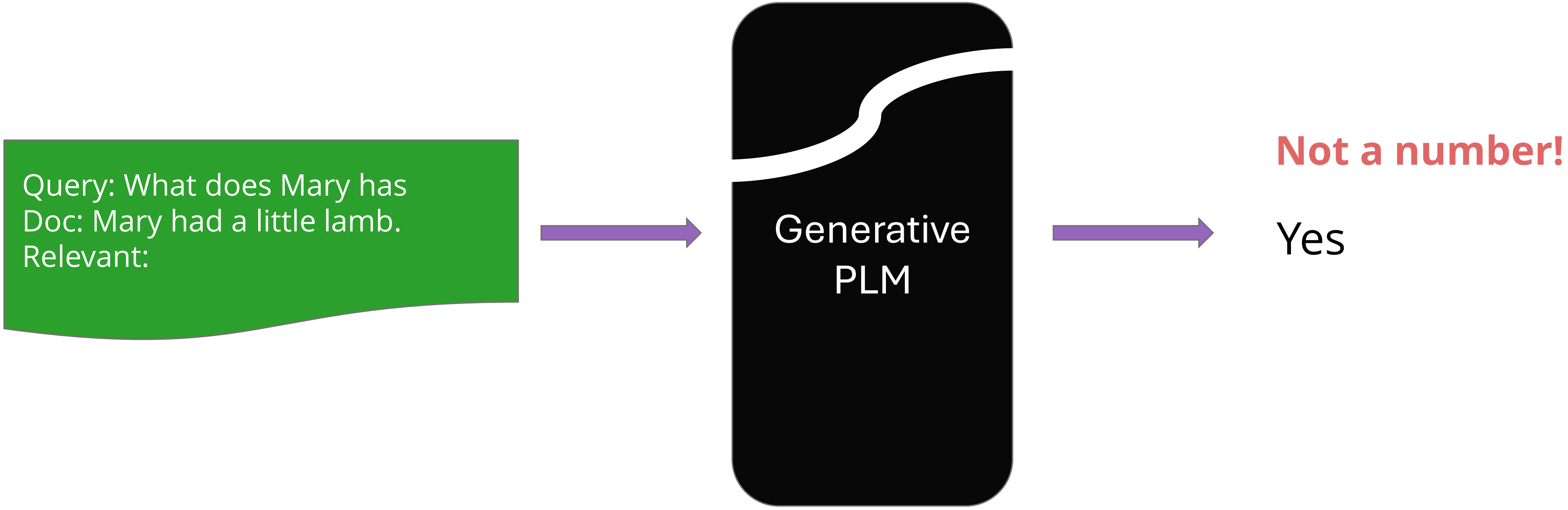
One **Sparse** Vector
Per Sequence
e.g., SPLADE

Multiple Dense Vectors
Per Sequence
e.g., ColBERT

Joint Encoder
e.g., monoBERT

Cross-Encoder

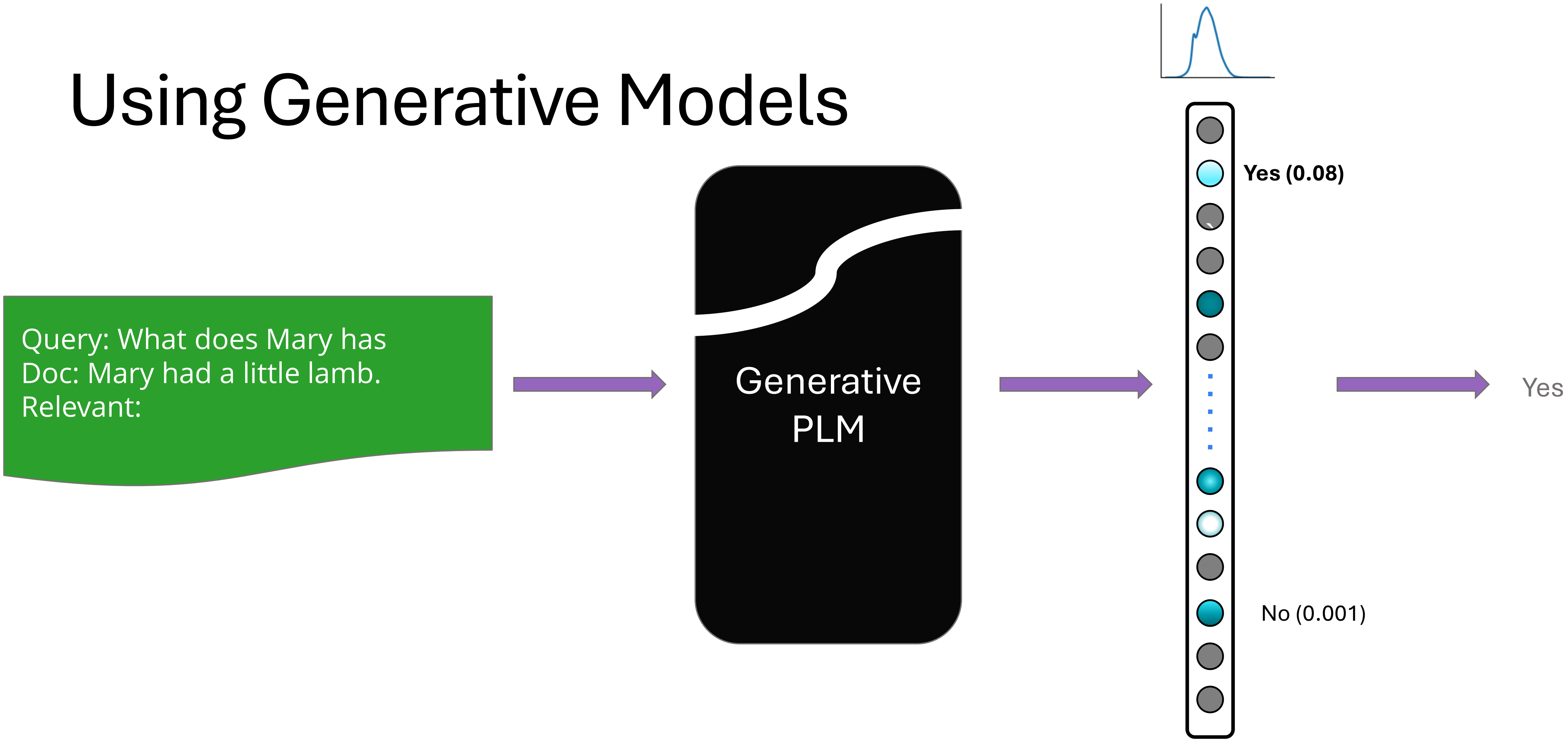
Using Generative Models



e.g, T5

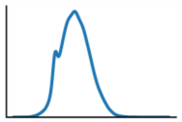
Pradeep, Ronak, Rodrigo Nogueira, and Jimmy Lin. "The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models." arXiv preprint arXiv:2101.05667 (2021).

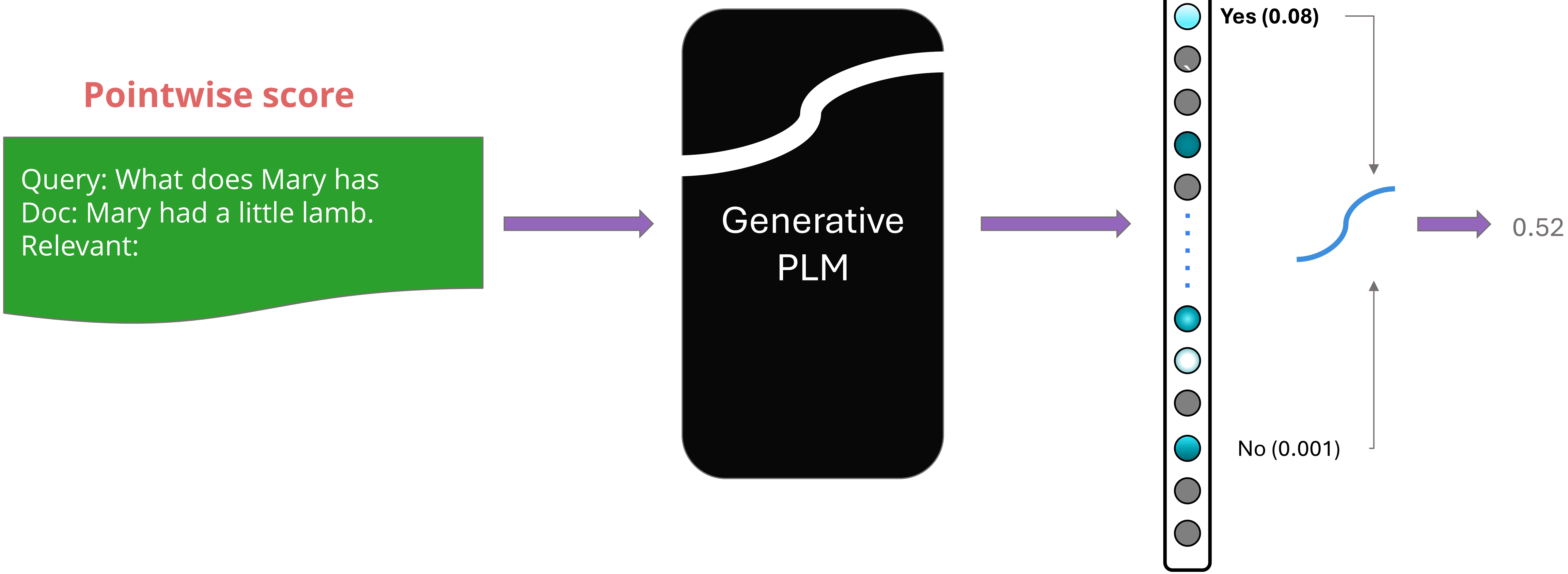
Using Generative Models



Pradeep, Ronak, Rodrigo Nogueira, and Jimmy Lin. "The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models." arXiv preprint arXiv:2101.05667 (2021).

Using Generative Models



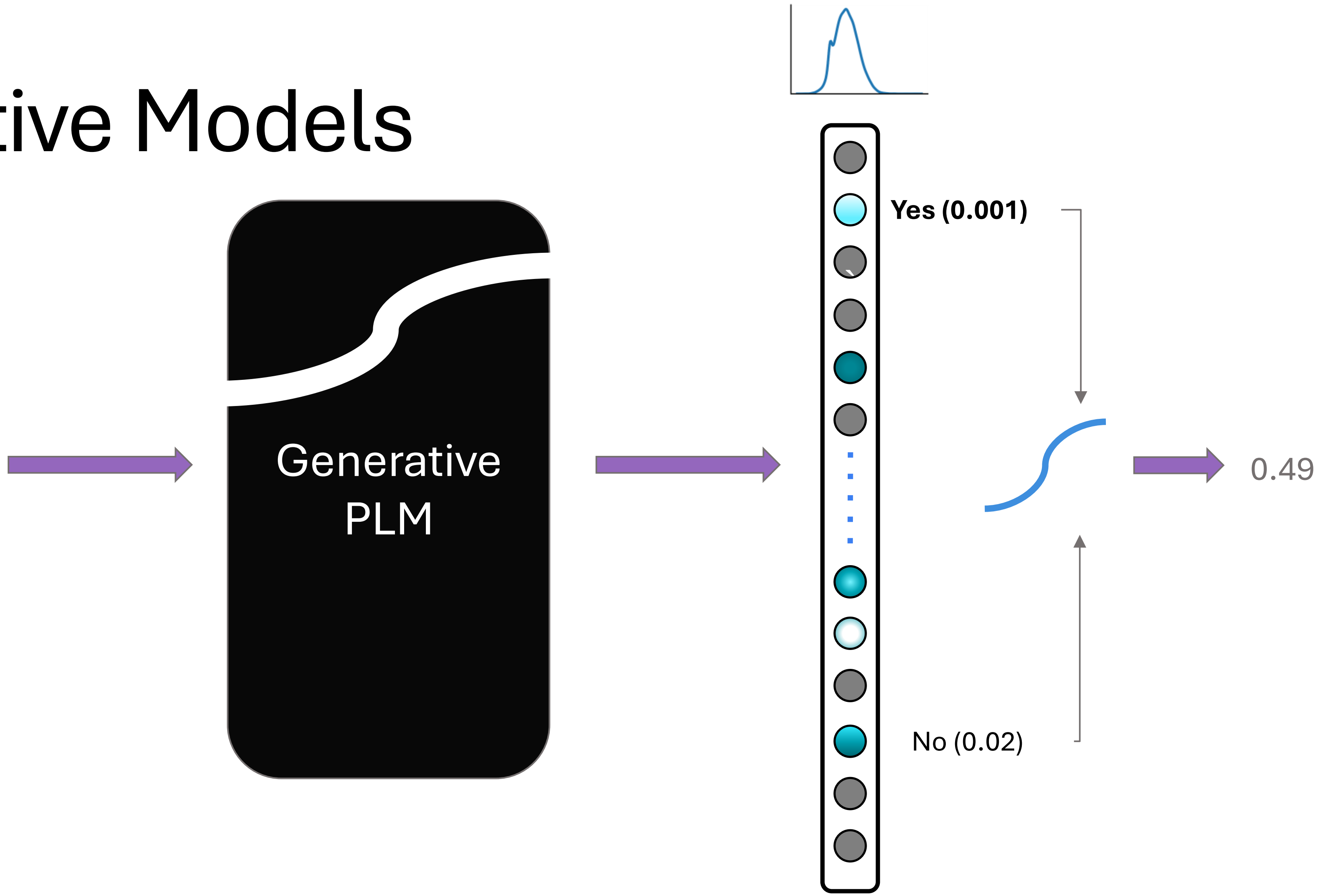


Pradeep, Ronak, Rodrigo Nogueira, and Jimmy Lin. "The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models." arXiv preprint arXiv:2101.05667 (2021).

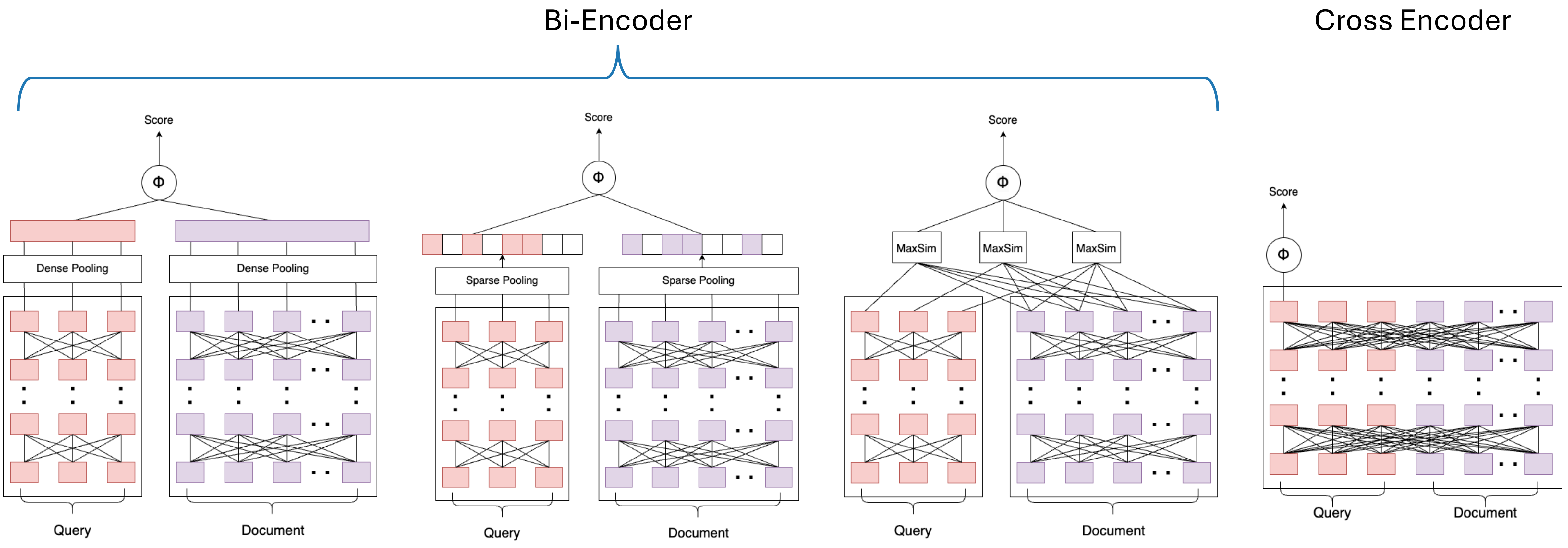
Using Generative Models

Pairwise score

Query: What does Mary has
Doc0: JHU is in Baltimore
Doc1: Mary had a little lamb.
Relevant:



Pradeep, Ronak, Rodrigo Nogueira, and Jimmy Lin. "The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models." arXiv preprint arXiv:2101.05667 (2021).

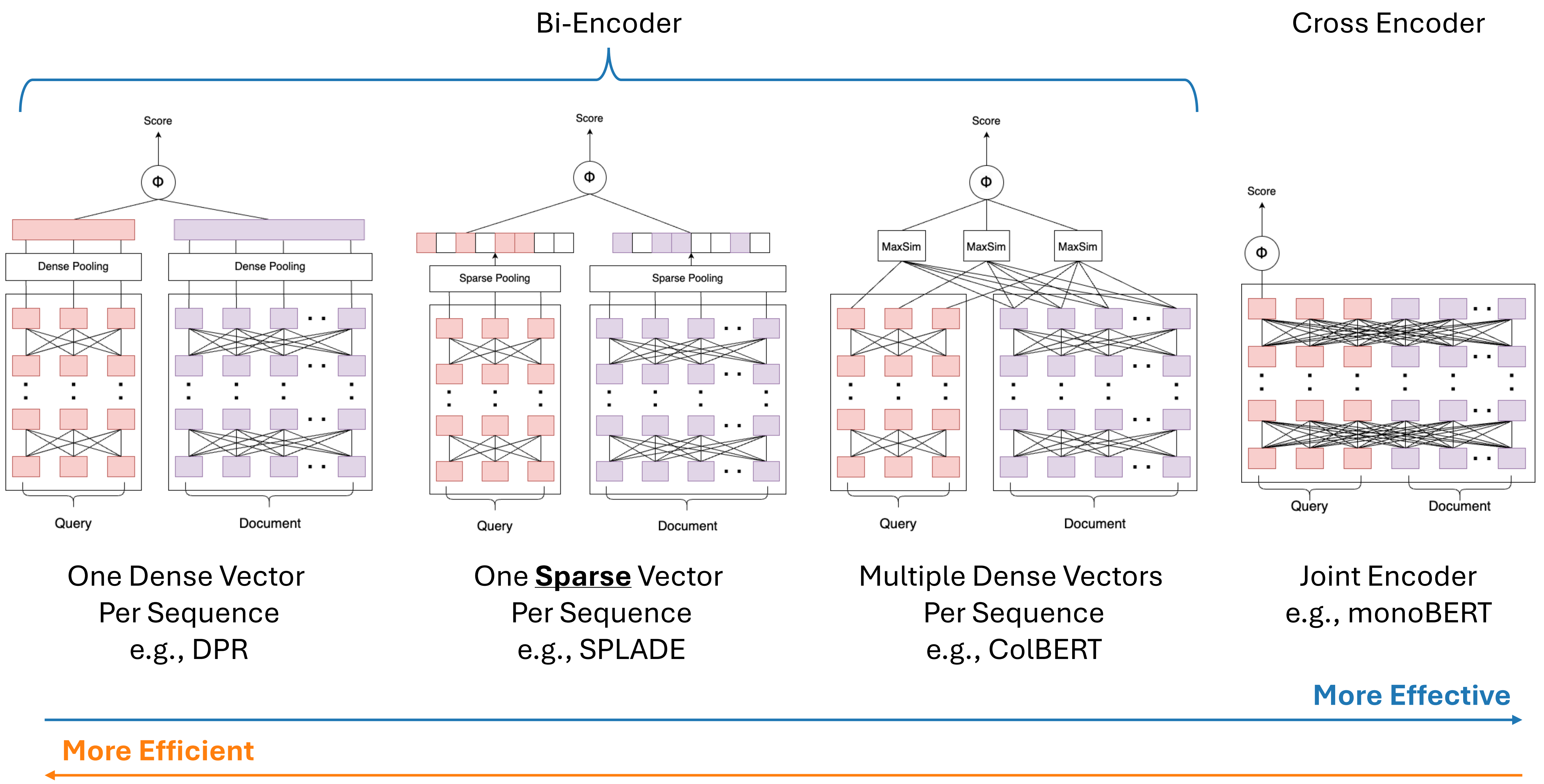


One Dense Vector
Per Sequence
e.g., DPR

One **Sparse** Vector
Per Sequence
e.g., SPLADE

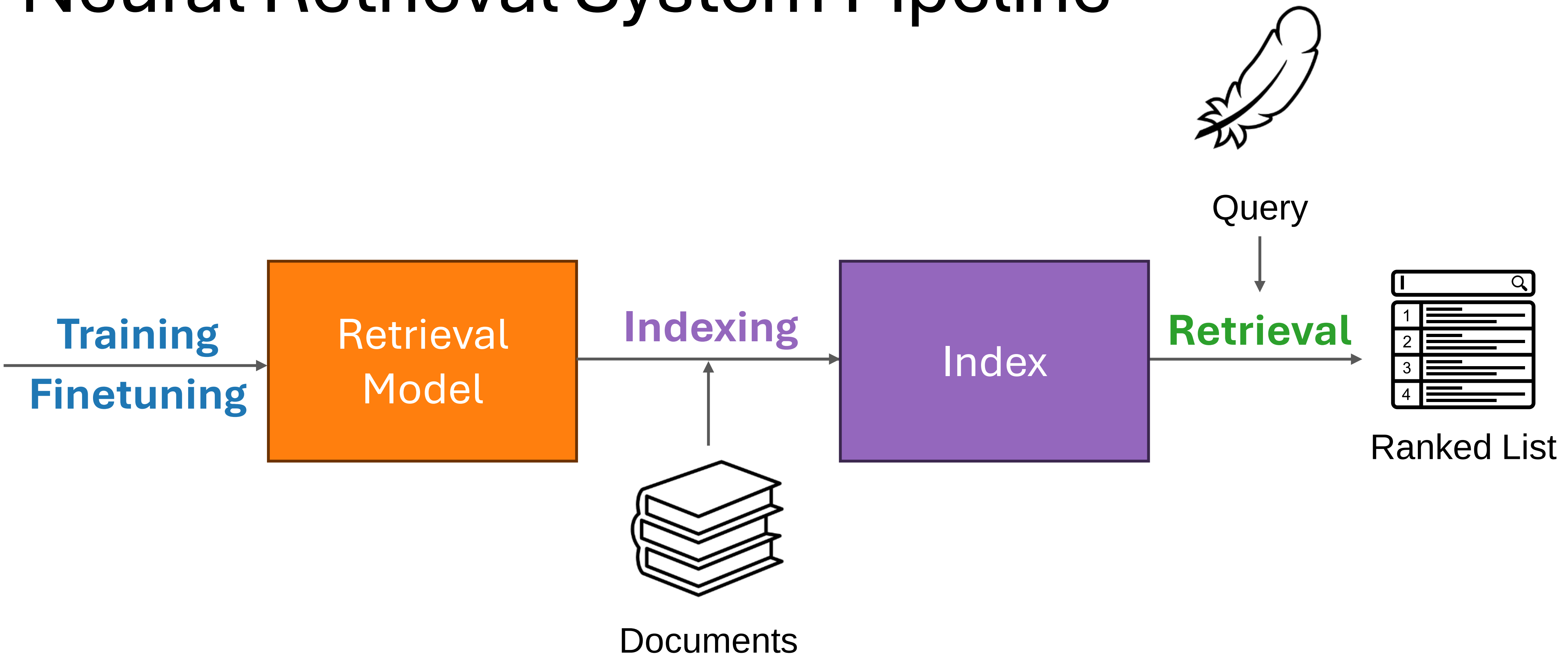
Multiple Dense Vectors
Per Sequence
e.g., ColBERT

Joint Encoder
e.g., monoBERT

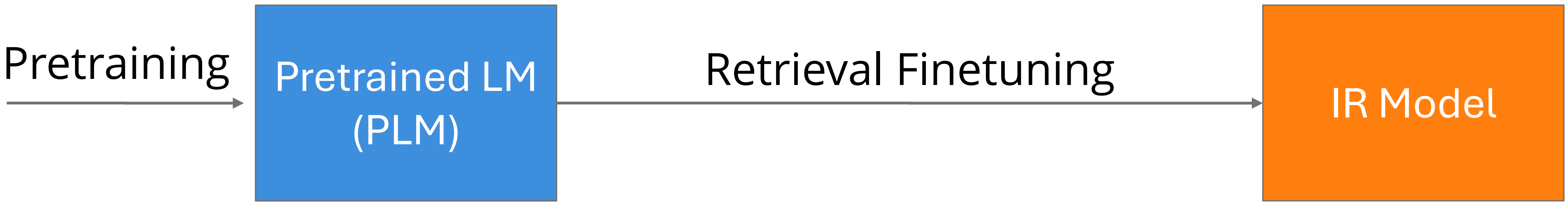

Bi-Encoder
Cross Encoder
Score
Φ
Dense Pooling
Dense Pooling
Query
Document
Score
Φ
Sparse Pooling
Sparse Pooling
Query
Document
Score
Φ
MaxSim
MaxSim
MaxSim
Query
Document
Score
Φ
Query
Document
One Dense Vector
Per Sequence
e.g., DPR
One Sparse Vector
Per Sequence
e.g., SPLADE
Multiple Dense Vectors
Per Sequence
e.g., ColBERT
Joint Encoder
e.g., monoBERT
More Effective
More Efficient

Retrieve-and-Rerank System Combinations

Neural Retrieval System Pipeline



PLM to IR Model



- Align the representation
- Model “relevancy”

Evaluation

Which system is better?

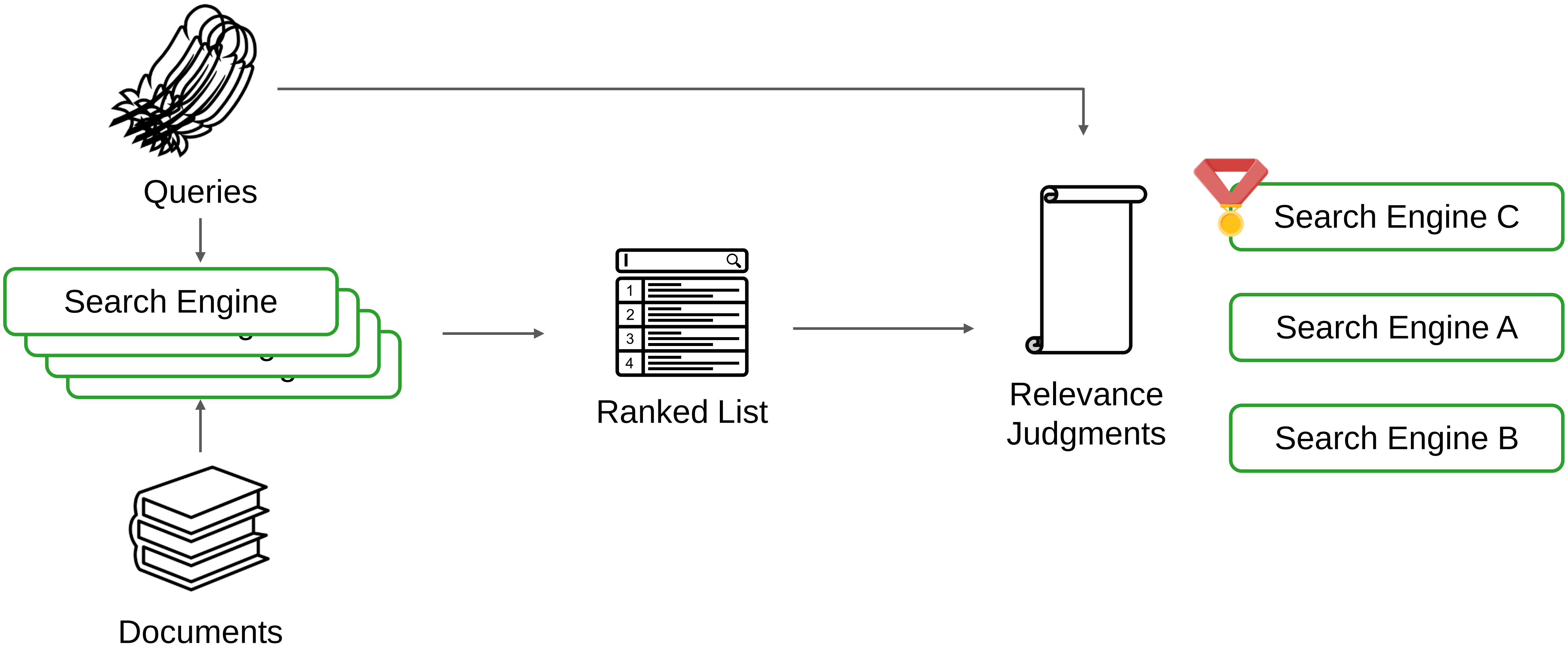
What is Information Retrieval?

(relevant)

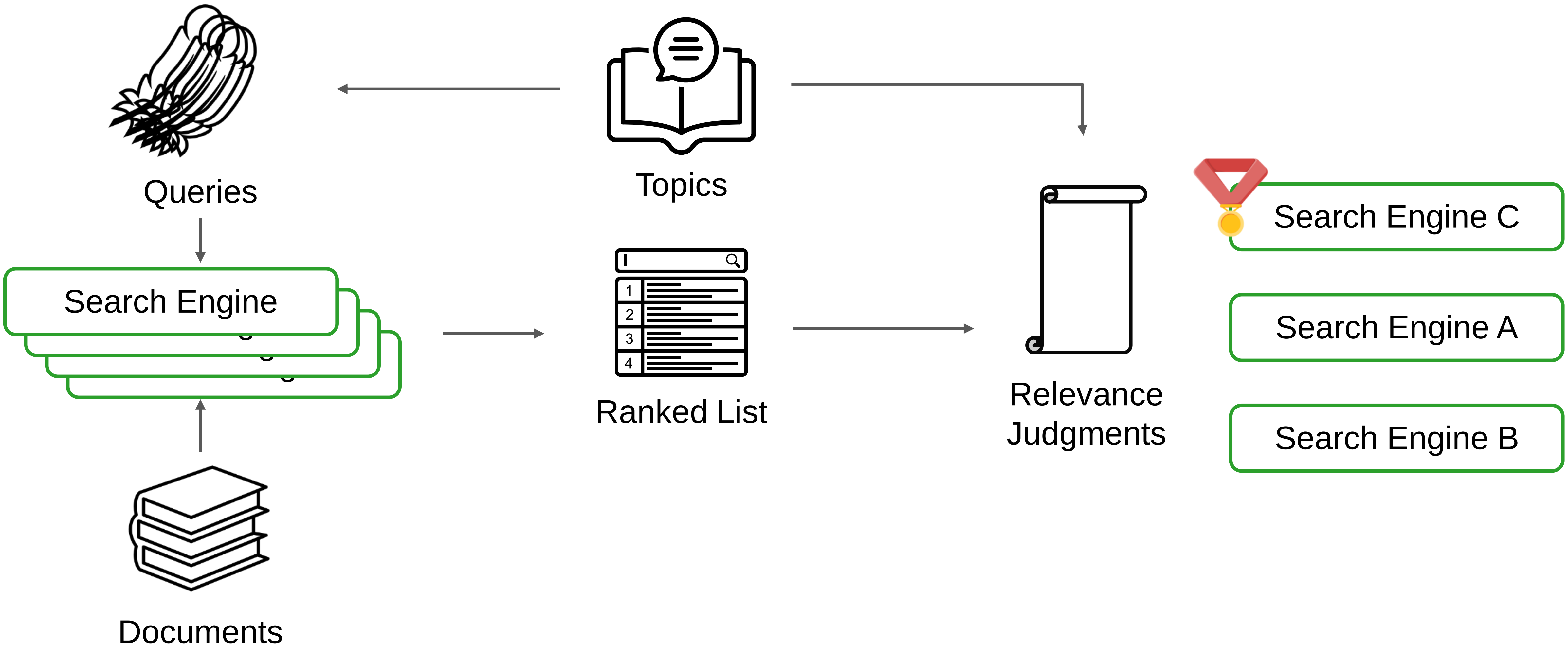
Retrieve information from a storage based on user's information need

Which system retrieve more relevant information?

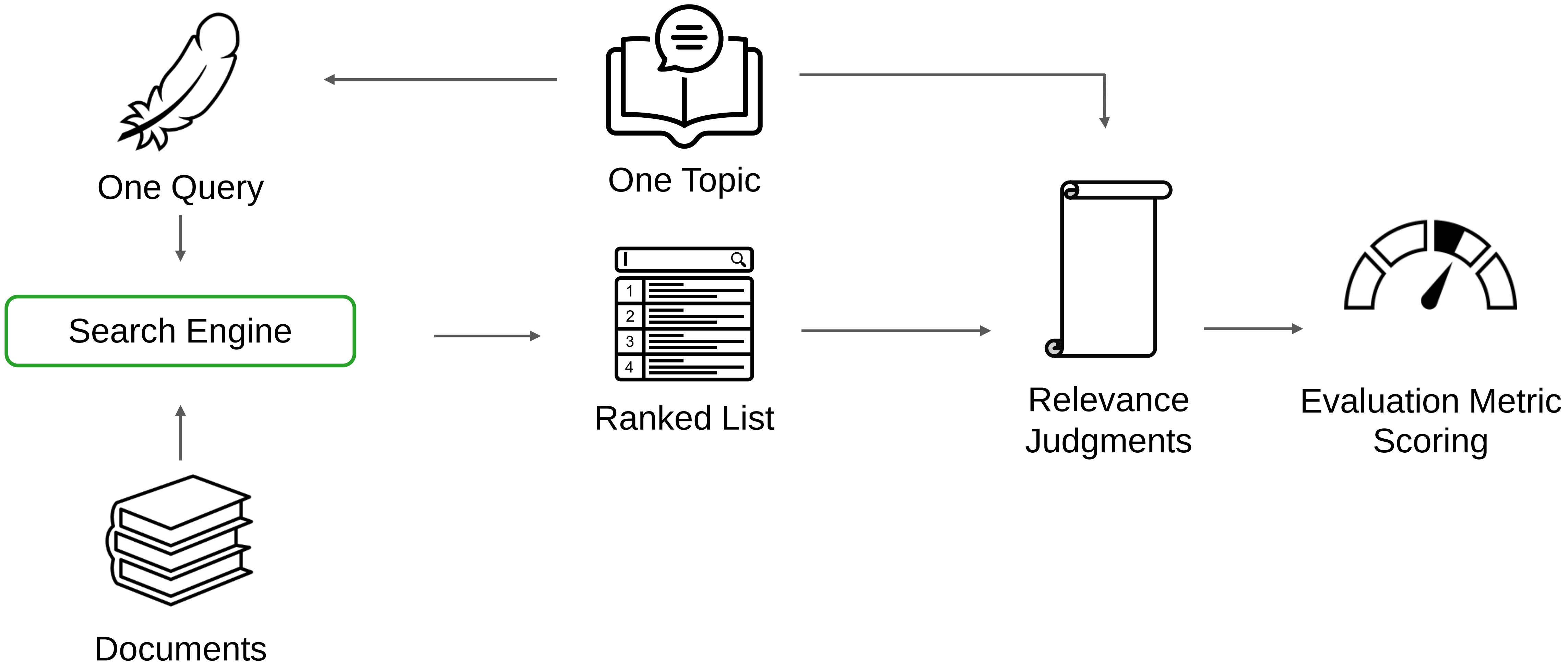
Cranfield Paradigm Evaluation



Cranfield Paradigm Evaluation



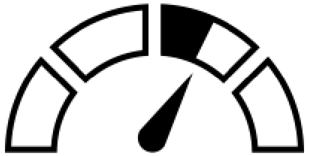
Cranfield Paradigm Evaluation



Differences

- Topics vs Queries
  - Clear intent vs an expression of such intent
- Relevance Judgements vs Labels
  - Opinion vs “fact”
- Ranked retrieval metrics
  - Measuring the quality/effectiveness of a ranked list

IR Metrics

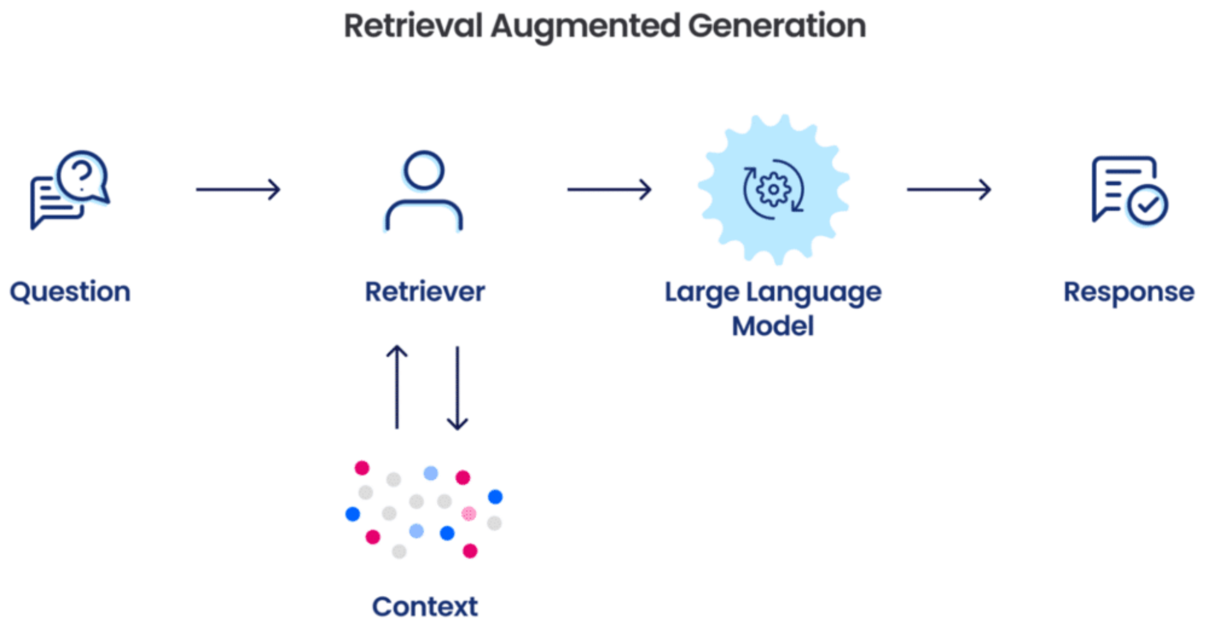


- Effective Metrics
  - Mean Average Precision
  - Normalized Discounted Cumulative Gain
  - Recall@k
- Efficiency Metrics
  - Indexing time
  - Index disk space
  - Query latency (average search time per query)

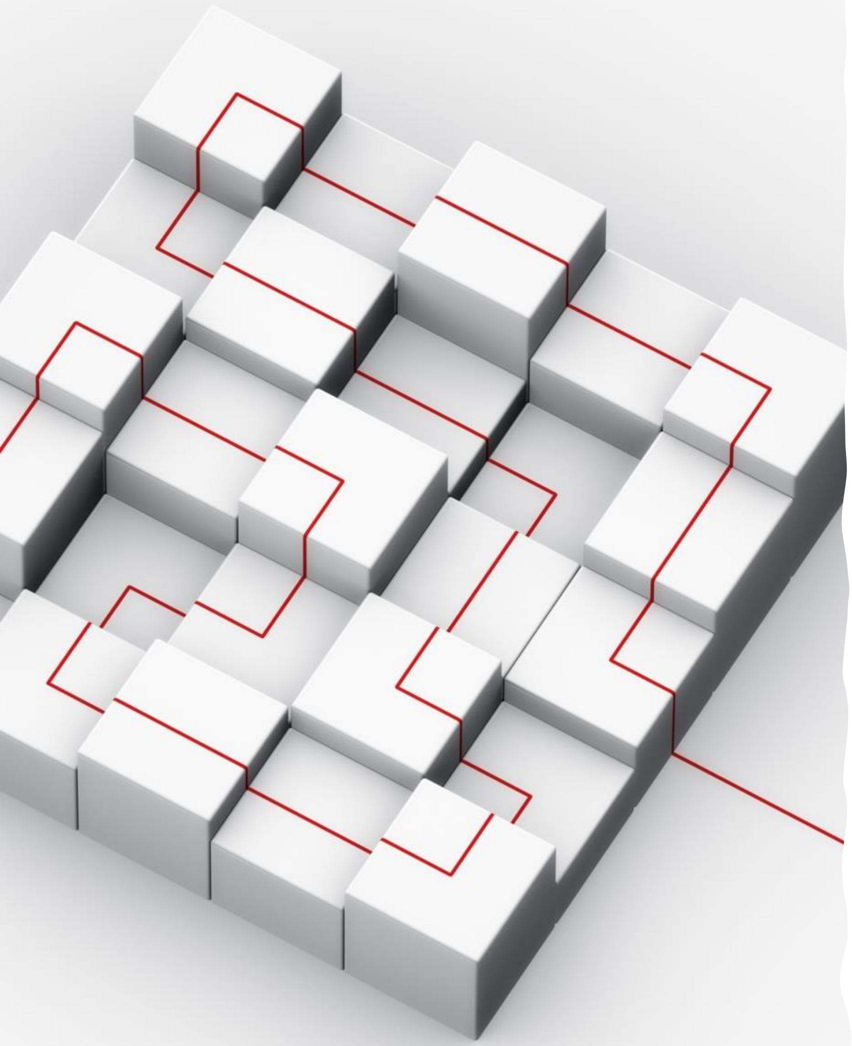
State of IR Research

Retrieval-Augmented Generation

- Is everything a RAG problem?
- What is the right retrieval model/system for RAG?
- IR going away?

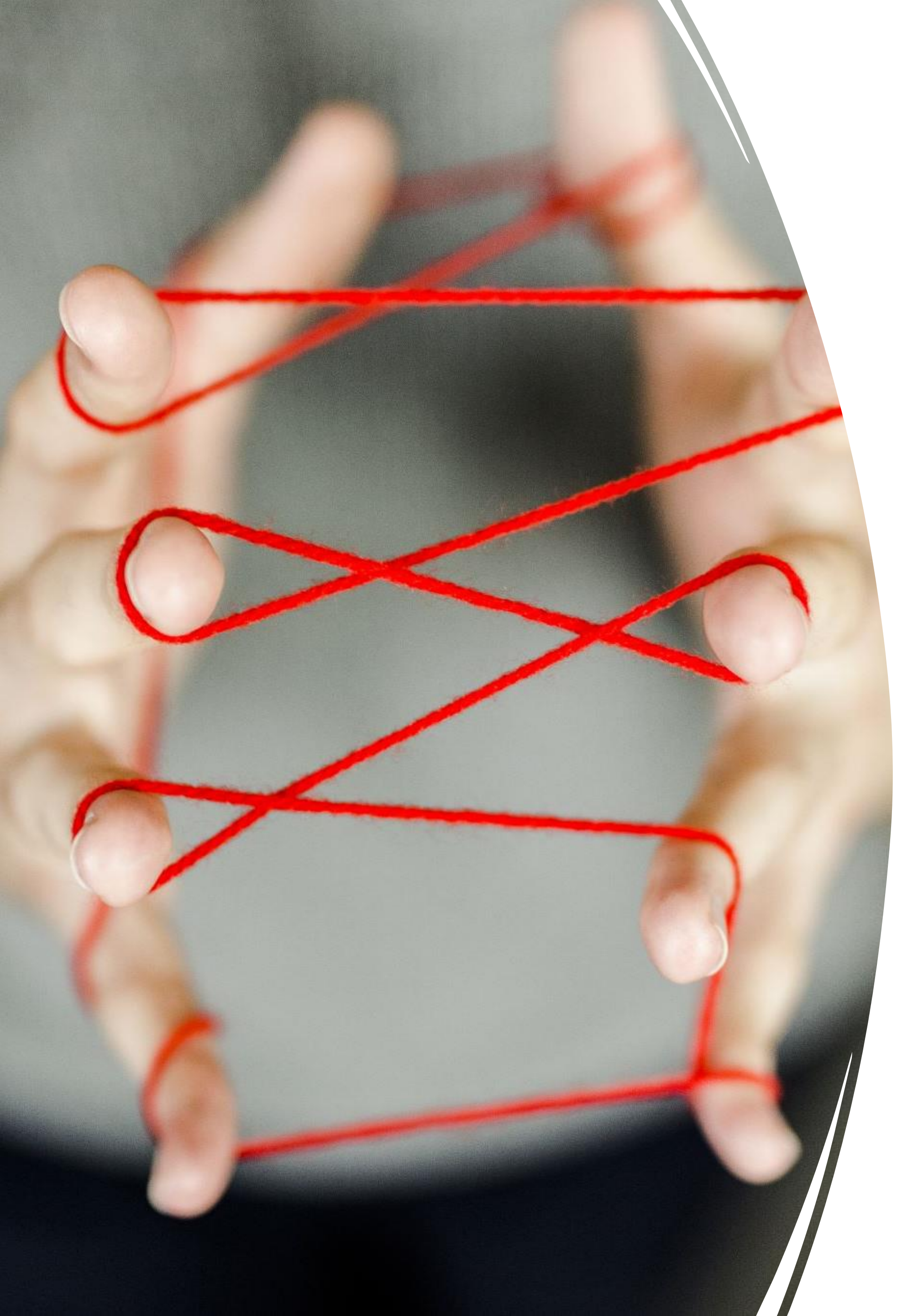


https://snorkel.ai/which-is-better-retrieval-augmentation-rag-or-fine-tuning-both/



Better Retrieval Models

- More effective
  - Better/larger neural models
  - Better architecture?
  - Under harder setup, e.g., scholar search, multilingual, cross-modal, etc
- More efficient
  - Faster at query time
  - Less resource footprint, e.g., memory, storage, compute, etc
- Other qualities
  - Fairness, diversity, etc



Other Retrieval Problems

- Conversational
  - Guessing intent, finding the “right” information to serve
- Iterative/interactive/human-in-the-loop
  - Rounds of interactions
- Generative
  - Returning a piece of text

Evaluation

- What to measure
  - and when would it fail
- How to measure
  - Generative text? Citations?
- “Better” evaluation collection
  - Not necessarily larger